EconomEtica

Centro interuniversitario per l'etica economica
e la responsabilità sociale di impresa
promosso dalla Fondazione Italiana Accenture

N.19 March 2010 - Revision Aug 2010

Giacomo Degli Antoni
Lorenzo Sacconi

Modeling Cognitive Social Capital and Corporate Social Responsibility (CSR) as Preconditions for Sustainable Networks of Relations

Working papers

Modeling Cognitive Social Capital and Corporate Social Responsibility (CSR) as Preconditions for Sustainable Networks of Relations

Giacomo Degli Antoni[♣] Lorenzo Sacconi[♦]

August 2010

Abstract

The paper studies the relationship between social capital (SC) and Corporate Social Responsibility (CSR) by investigating the idea of a virtuous circle between the level of SC and the implementation of CSR standard of behaviour that favours the creation of cooperative networks between the firm and all its stakeholders by promoting the spread of social norms of trust and cooperation.

Multidimensionality of social capital (Uphoff 1999, Paldam 2000) is accounted in terms of cognitive and structural SC. The first refers to dispositional characters of agents that affect their propensity to behave in conformity with social norms, whereas the latter consists of social networks connecting agents. With regard to the concept of CSR, we adopt a contractarian approach and consider CSR as an extended model of corporate governance, based on fiduciary duties owed to all the firm's stakeholders. Among stakeholders, we distinguish between strong and weak stakeholders. Both of them are locked into a relation with the firm by specific investments. While, however, cooperation with strong stakeholders is a long run equilibrium for the firm, on the contrary, in the relations with weak stakeholders the firm face material incentives to defect from cooperation with them.

By joint use of the tools of network analysis and psychological game theory, the paper shows the role of cognitive SC and CSR in promoting the emergence of cooperative networks between the firm and all its stakeholders (structural SC). In particular, (a) the level of cognitive SC, in terms of community or society-wide disposition to comply with fair social norms, plays a key role in providing opportunities for the firm to agree (with strong stakeholders) on CSR principles and hence to induce incentives to comply with them. (b) The explicit agreement on CSR principles and norms engenders cognitive social capital on its own. It does so by creating room for conformist preferences that exploit beliefs of mutual conformity and dispositions to conform. Moreover, the agreement on CSR principles by itself positively affects beliefs about reciprocal conformity on the part of the firm and its strong stakeholders. (c) The level of cognitive social capital (both beliefs and dispositions) and the decision to adopt CSR principles generate structural social capital understood as long-term cooperative relationships between the firm and its stakeholders, even though, on considering the material payoffs characterizing the single relationships, the firm would have no incentive to cooperate with weak stakeholders. Alongside the notion of sub game prefect equilibria and credible threats, we show that strong stakeholders endowed with high cognitive social capital, have an incentive in punishing the firm if it is not cooperative with weak stakeholders. The sanction may induce the firm to cooperate with weak stakeholders as well, and it generates cooperative networks that would not be sustainable without the power of the sanction.

Keywords: Social capital, Corporate Social Responsibility, Social norms, Network, Cooperation, Trust.

JEL Classification: A13; D23; L21; M14; Z10

To be published in L. Sacconi and G. Degli Antoni (eds), *Social Capital, Corporate Social Responsibility, Economic Behavior and Performance* (Basingstoke: Palgrave MacMillan), forthcoming.

♣ Department of Sociology and Social Research, University of Milano - Bicocca and EconomEtica, interuniversity centre of research, University of Milano-Bicocca. Email: giacomo.degliantoni@unimib.it.

♦ Department of Economics, University of Trento and EconomEtica, interuniversity centre of research, University of Milano-Bicocca. Email: lorenzo.sacconi@economia.unitn.it

1 Introduction

1.1. Subject and aim

In recent years, increasing attention has been paid to trust, trustworthiness and social norms of reciprocity and cooperation as key factors in socio-economic development. Even though from different perspectives, both the concept of social capital and the notion of corporate social responsibility refer to these elements.

Since the seminal work by Putnam, Leonardi and Nanetti (1993) focusing on the effects of social capital (hereafter also SC) on economic and government performance, the concept of SC has been widely used to analyse how interpersonal relations affect economic activity by favouring cooperation. Many definitions of social capital have been proposed, and two principal approaches to this concept may be identified. On the one hand, social capital is defined in terms of generalised trust, civic norms, beliefs and dispositions which affect the propensity to cooperate (e.g. Putnam et al., 1993; Knack and Keefer, 1997). On the other hand, social capital is defined in terms of cooperative networks among agents (e.g. Coleman, 1988; Lin, 2001; Burt, 2002). Many approaches are also taken to the notion of corporate social responsibility (hereafter also CSR). In particular, if we consider the stakeholder approach (Freeman 1984 and 2000; Freeman and Evan, 1990) or the contractarian approach to CSR (Sacconi 2004, 2006, 2007a and 2007b), relational aspects in terms of trust, trustworthiness, beliefs and dispositions to cooperate seem to be fundamental in promoting the coordination processes between the firm and its stakeholders that are essential to implement CSR practices.[1] Even though SC and CSR seem to share several features, their relationship has not yet been analysed in depth.

In this paper we model the relationship between the firm and its stakeholders and show analytically how (cognitive) social capital and corporate social responsibility generate (structural) social capital.

[1] Relational elements concerning the relationship between the firm and its stakeholders are indubitably less important if we look at other CSR approaches. Neither Friedman (1977) nor Jensen (2001), for example, give much space to explicit consideration of the stakeholder's interests by the owners of firms. The idea of Friedman is that the only social responsibility of a firm is to make profits while respecting the rules, which means without breaking the law. Jensen's contention is that in the long term maximization of the shareholder value is the best way to satisfy also the stakeholders' interests that the multi-stakeholder approach to CSR wants to protect.

1.2 Social capital

Taking into account the multi-dimensional character of SC (e.g. Paldam, 2000), and starting from the distinction drawn by Uphoff (1999), we consider a cognitive and a structural dimension of the concept. In our approach, the former dimension essentially refers to the dispositional characters of agents that affect their propensity to behave in different ways. The latter refers to social networks that connect agents. More specifically, we approach the idea of cognitive social capital by focusing on trustworthy attitudes grounded on preferences for social norm compliance, in turn based on reciprocal beliefs and more basic dispositions to conformity. Reciprocal beliefs (in the behaviour of others) depend on the behaviour that others have already exhibited in the past but can be generated (or reinforced) by ethical commitments undertaken by them (for example, if agents subscribe to an agreement on an ideal principle). Dispositions stem principally from more basic cultural traits in the community where agents live; but they also depend on micro elements (e.g. genetic and psychological factors). Both beliefs and dispositions can promote (or, obviously, reduce) trust and propensity to cooperate. Structural social capital is constituted by cooperative linkages among agents. We consider four main factors able to promote the creation of structural social capital (three pertaining to the cognitive dimension of social capital, the fourth to the structure of interaction): (i) reciprocal beliefs that others will cooperate, (ii) disposition to cooperate, (iii) agreements on social norms and principles that may activate reciprocal beliefs and dispositions and translate them into motives to act (this is the point where the logical connection with CSR will become stringent) and (iv) the existence of credible sanctions against the agents that decide not to cooperate.[2] Our definitions of structural and cognitive social capital differ from those proposed by Uphoff. However, they share some essential characteristics with them. In regard to the structural definition, both our approach and that adopted by Uphoff include in this dimension the networks that contribute to cooperation. In regard to the cognitive dimension, Uphoff's approach states that this category "derives from mental processes and resulting ideas, reinforced by culture and ideology, specifically norms, values, attitudes, and beliefs that contribute cooperative behavior" (Uphoff, 1999, p.218). We refer to cognitive social capital by

[2] See Sacconi and Degli Antoni (2009) for a deeper discussion of these notions of cognitive and structural social capital.

focusing on beliefs and dispositions, and we show how they affect the propensity of agents to share ethical principles of cooperation.

1.3 Corporate social responsibility

We take a contractarian approach to corporate social responsibility and define it as a 'model of extended corporate governance whereby those who run a firm (entrepreneurs, directors and managers) have responsibilities that range from fulfilment of their fiduciary duties[3] towards the owners to fulfilment of analogous fiduciary duties towards all the firm's stakeholders' (Sacconi, 2006). The definition of CSR in terms of extended responsibility towards all the stakeholders of the firm is rooted in neo-institutional theory (Williamson, 1975 and 1986; Grossman and Hart, 1986; Hart and Moore, 1990; Hart, 1995; Hansmann, 1996). According to this approach, the firm is an institutional form of 'unified transactions governance' aimed at remedying imperfections in the contracts that regulate exchange relations among subjects endowed with diverse assets (capital, labour, instrumental goods, and so on) that may generate a surplus if put together. The incompleteness of contracts that should regulate the agreements on the investment to be made by each agent, and on how the surplus is to be divided among them, reduces the incentive of subjects to invest at an optimal level. The firm responds to this problem by bringing the various transactions under the control of a hierarchical authority which owns the firm and is entitled by its ownership to make decisions on the contingencies that were not *ex ante* contractible.[4] This party is thus safeguarded against opportunism by the other stakeholders. Nevertheless, this configuration generates a risk for the other parties, which are vulnerable to an abuse of authority (Sacconi, 1999, 2000 and 2006). Many non-controlling stakeholders will *ex ante* be discouraged from investing at an optimal level, while *ex post* they will resort to conflicting or disloyal behaviour (typically possible when information asymmetry is inherent in the execution of some subordinate activity), in the belief that they are being subjected to the abuse of authority. Therefore, the optimal level of investment cannot be achieved and a second-best solution arises. This result, which approximates social efficiency, is always connected with governance solutions based on the allocation of property rights to a single party.

[3] On the concept of fiduciary duty see Flannigan (1989) and Sacconi (2006).

According to the contractarian approach, this problem can be overcome if CSR is viewed as 'extended governance' (Sacconi, 2000 and 2006). The firm's legitimacy deficit is remedied if the residual control right is associated with further fiduciary duties of the controlling stakeholder towards the non-controlling ones faced with the risk of abuse of authority. The firm must be grounded on a rational agreement (the constitutional contract of the firm) between those who run it (entrepreneurs, directors and managers) and the non-controlling parts (Sacconi, 2006). It is the constitutional contract of the firm which determines

- that authority is delegated to the stakeholder most efficient in performing governance functions;

- the fiduciary duties of this party towards the non-controlling stakeholders.

Once the social contract of the firm has been defined, the firm must develop a reputation in order to convince all the non-controlling stakeholders that it will respect the duties stipulated in the contract. The problem with creating reputation is that the firm and its stakeholders are characterized by settings in which information or knowledge about the action of the firm is incomplete or highly asymmetric.[5] Because of incomplete information, the stakeholders cannot verify whether the firm has actually behaved according to the fiduciary duties defined in the social contract and, consequently, the firm cannot develop a reputation. In order to do so, it must adopt an explicitly announced standard (a CSR standard) that sets out general principles and whose contents are such to elicit stakeholder consensus, as well as explicit commitments to comply with principles and rules known *ex ante* by stakeholders.[6]

1.4 Weak and strong stakeholder

Finally, with respect to the term 'stakeholder', which denotes individuals or groups with a major stake in the running of the firm and that are able to influence it significantly (Freeman and McVea, 2002), we accept the distinction between

[4] The decision about the party that must have the residual right of control may depend on various factors - e.g. a comparative analysis of the control costs of the various stakeholders: see Sacconi (2006) for a deeper explanation.
[5] For a deeper explanation of this theory of reputation under unforeseen contingencies see Sacconi (2000 and 2004).
[6] For the design of a CSR management standard that corresponds to the features now defined: Sacconi DeColle Baldin (2003) and Clarkson Centre for Business Ethics (2002).

stakeholders in the strict or in the broad sense. The former are stakeholders who have an interest at stake because they have made specific investments in the firm (i.e. investments that may significantly increase the total value generated by the firm and that are made in relation to a specific firm and not any other). Stakeholders in the broad sense are stakeholders connected to the firm because they *undergo* the 'external effects' of the transactions performed by it, even if they do not directly participate in those transactions. With respect to this classification we introduce, within the category of stakeholders in a strict sense, the original distinction between strong and weak stakeholders. Strong and weak stakeholders are distinguished by the consequences that the breaking-off of the relationship with the firm produces both on the stakeholder and on the firm.

a) *Strong stakeholder*. The difference between the discounted payoff that strong stakeholders and firms obtain by cooperating forever and by defecting at the first stage (and never cooperating again) is positive. Strong stakeholders are stakeholders in the strict sense that bring strategic assets into the firm. They are, for example, highly skilled workers or institutional investors.

b) *Weak stakeholder*. Weak stakeholders would like to cooperate forever with the firm, but the discounted payoff that the firm obtains by cooperating forever with them is lower than the payoff it obtains by defecting at the first stage and never cooperating again. Weak stakeholders are stakeholders in the strict sense who do not bring strategic assets into the firm. They are, for example, ordinary investors, unskilled workers or unskilled contractors.

1.5 Main results and outline of the chapter

Considering the notions of cognitive and structural SC, a contractarian approach to CSR and the distinction between strong and weak stakeholders, we develop a model that yields three main results.

1. The level of cognitive SC, in terms of generic community or society-wide disposition to comply with fair social norms, plays a key role in providing opportunities for the firm to agree (with strong stakeholders) on CSR principles of fairness and hence to induce incentives to comply with them with respect to all the stakeholders, especially the weak ones.

2. The explicit agreement on CSR principles and norms engenders cognitive social capital on its own. It does so by creating room for conformist preferences that exploit beliefs of mutual conformity and dispositions to conform by converting them into specific reasons to comply with an agreed principle of CSR. Moreover, the agreement on CSR principles of fairness by itself (through framing effects and default reasoning) positively affects beliefs about reciprocal conformity on the part of the firm and its strong stakeholders.

3. The level of cognitive social capital (both beliefs and dispositions) and the decision to adopt CSR principles and norms (that translates the former into conformist preferences) generate structural social capital understood as long-term cooperative relationships between the firm and its stakeholders, even though, on considering the material payoffs characterizing the single relationships, the firm would have no incentive to cooperate with weak stakeholders. We show that strong stakeholders endowed with high cognitive social capital, which start cooperating with a firm that adopts a CSR standard, have an interest in punishing the firm if it is not cooperative with weak stakeholders. The sanction may induce the firm to cooperate with weak stakeholders as well, and it generates cooperative networks that would not be sustainable without the power of the sanction.

The second section presents the analytical framework used to study the networks of relations between firms and stakeholders. It analyzes these relations by considering Prisoners' Dilemmas (with respect to the relationship between the firm and weak stakeholders) and an enlarged version of the Trust Game (relationship between the firm and strong stakeholders), also illustrating a basic flaw in this literature on social capital. The third section considers the possibility that agents are not motivated exclusively by material payoffs (the idea of conformist preferences is introduced) and reinterprets the relationship between the firm and its strong stakeholders by introducing a psychological game with its psychological payoffs and equilibria. This section illustrates the role of cognitive social capital in affecting the psychological payoff of the firm and of strong stakeholders. Section four shows how cognitive social capital (in terms of disposition), agreed CSR principles, and learning from iterated games played in the network affect the strong stakeholder's strategy in interacting with the firm. Discussed in particular is

the effect of CSR and of the firm's behaviour in repeated games with its weak stakeholders on strong stakeholders' belief formation and strategy. It is argued that cooperation in the network is supported by cognitive social capital. The fifth section analytically presents the mechanism behind the formation of firm's and strong stakeholders' beliefs and the strategies determined by how iterated games involving the firm and all its stakeholders in the network are played. Thus repeated strategies are defined that induce cooperation and the endogenous sanctioning of 'defection' and 'unfair behaviour'. Section six verifies that the strategies inducing cooperation in all the games the firm plays with its stakeholders satisfy a condition of sustainability and stability in the psychological game played by the firm and its strong stakeholders, this being seen as a stage sub-game in the entire iterated interaction among all the participants in the network. Herein resides the paper's main result: the demonstration that, due to conformist preference and psychological payoffs (i.e. the way in which the model depicts the players' cognitive social capital) cooperative behaviour throughout the entire network (namely the emergence of structural social capital) is a sub-game perfect equilibrium due to the stage-game equilibria of the psychological game wherein strong stakeholders have the proper incentive to punish the firm's deviations from a strategy of multilateral cooperation. Section seven identifies and verifies the conditions guaranteeing that the multilateral cooperative strategy played by the firm in the repeated games with each of its stakeholder satisfies the condition for the existence of repeated games Nash equilibria. In accordance with standard treatments of repeated games, it is shown that, when cognitive social capital is sufficiently high and beliefs are coherent with the cooperative equilibrium in the psychological game, for reasonable values of the firm's discount factors δ, the firm will cooperate also with weak stakeholders in order to continue its cooperation with strong stakeholders. Section eight concludes.

2. A relational network involving the firm and its (strong and weak) stakeholders

2.1 The analytical framework

We will analyse the relational networks between firms and stakeholders by using the analytical framework suggested by Lippert and Spagnolo (2009) (hereafter L&S), which is summarized here for the reader's convenience (see also Lippert, 2010, *infra*). L&S study relational networks in order to investigate the power of sanctions and networks' equilibrium conditions under different configurations and information transmission

technologies. Consider a set $N = \{1,...,n\}$ of infinitely lived agents $i \in N$. The agents can interact in pairs according to a connection structure C of two element subsets of N. C_i is the set of connections that characterizes agent i. In each period t, the agents that are connected play a Prisoner's Dilemma (PD) with payoffs given by the matrix of Figure. 1. The payoff structure is: $l_{i,j}<d_{i,j}<c_{i,j}<w_{i,j}$ and $l_{i,j} + w_{i,j} <2c_{i,j}$, $\forall\ i, j \in N$, $i \neq j$ and the stage game is assumed to be constant over time. The payoffs imply the static Nash equilibrium ($D_{i,j}$, $D_{j,i}$). Agents are assumed to interact repeatedly; time is discrete; all agents are assumed to have a discount factor $\delta<1$;[7] agents are assumed to aim at maximizing their discounted utility.

Figure 1 Generalized form of the PDs played by pairs of players located at any adjacent mode of the network

		Agent j	
		C_{ji}	D_{ji}
Agent i	C_{ij}	$c_{i,j}$, $c_{j,i}$	$l_{i,j}$, $w_{j,i}$
	D_{ij}	$w_{i,j}$, $l_{j,i}$	$d_{i,j}$, $d_{j,i}$,

According to L&S's definition, two agents share a relation (R) if they repeatedly play (C_{ij},C_{ji}). Individual gains are defined by means of the following notation: g_{ij} is the net expected discounted gain of agent i from the relation with player j and it is the difference between the discounted payoff that agent i gets by playing ($C_{i,j}$, $C_{j,i}$) forever and defecting and starting to play the static Nash equilibrium ($D_{i,j}$, $D_{j,i}$) thereafter:

$$g_{ij} \equiv c_{i,j} - (1-\delta)w_{i,j} - \delta d_{i,j}$$

A relation of player i with player j in which $g_{ij}<0$ is called a 'deficient relation' for player i; a relation of player i with player j in which $g_{ij}\geq 0$ is called 'non-deficient' for player i; a relation between i and j is called 'mutual' iff $g_{ij}\geq 0$ and $g_{ji}\geq 0$; it is called 'unilateral' iff either $g_{ij}<0$ and $g_{ji}\geq 0$ or $g_{ij}\geq 0$ and $g_{ji} < 0$; and it is called 'bilaterally deficient' iff $g_{ij}<0$ and $g_{ji}<0$.

[7] Additive separability of agents' payoffs across interactions and across time is assumed for simplicity.

A graphical representation of the possible kinds of relations between *i* and *j* according to the value of g_{ij} is as follows:

- an incoming arrow to player *i* represents a non-deficient relation for player *i* (i.e $g_{ij} \geq 0$)
- an outgoing arrow from player *i* represent a deficient relation for player *i* (i.e $g_{ij} < 0$).

According to the above definition, Figures 2a) and b) depict unilateral relations. Figure 2c) depicts a mutual relation and Figure 2d) depicts a bilaterally deficient relation.

Lippert and Spagnolo (2009) start from this framework to analyse the sustainability of different network configurations under three information transmission mechanisms (Perfect Information Transmission; No Information Transmission; Network Information Transmission) and considering two types of multilateral strategy: multilateral grim trigger strategies and multilateral repentance strategies.

Figure 2 Graphical representation of relations

a) $j \longrightarrow i$ $\quad g_{ij} \geq 0$ and $g_{ji} < 0$

b) $j \longleftarrow i$ $\quad g_{ji} \geq 0$ and $g_{ij} < 0$

c) $j \longleftrightarrow i$ $\quad g_{ij} \geq 0$ and $g_{ji} \geq 0$

d) $j \quad i$ $\quad g_{ij} < 0$ and $g_{ji} < 0$

We focus our analysis on the situation under perfect information transmission considered by L&S. Under Perfect Information Transmission every player observes the actions taken by any other player in the network.[8] It can be shown that a sustainable

[8] We will slightly modify this assumption in our model.

strategy profile for the network is the adoption by every agent of the MG trigger strategy:

Every player $i \in N^s$

1. starts playing C_{ij} $\forall j \in R_i$,

2. continues playing C_{ij} $\forall j \in R_i$ as long as s/he observes C_{mn} $\forall m,n \in N^s$, and

3. reverts to D_{ij} $\forall j \in R_i$ forever otherwise.

The resulting relational network is sustainable if each player prefers to cooperate with all his/her neighbours rather than deviating from playing cooperatively with regard to any subgroup of them and facing retaliation from all neighbours. If a player decides to deviate from his/her relations with any subgroup of his/her neighbours, s/he faces retaliation from all neighbours and can thus just as well (and should optimally) deviate from all his/her relations. In terms of net gains from cooperation this result can be expressed as follows:

Under Perfect Information Transmission (I1), a relational network is sustainable if and only if $\sum_{j \in R_i} g_{ij} \geq 0 \forall i \in N^s$

The following Table 1 summarizes the basic notation used throughout the paper.

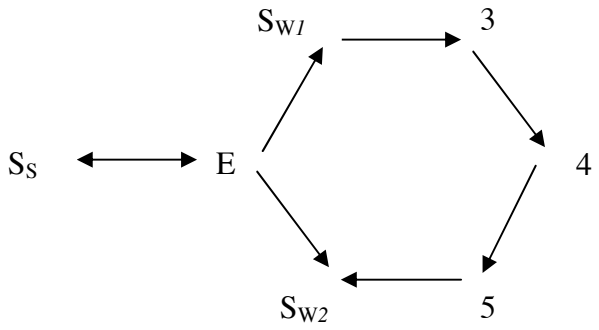
Table 1 Basic notation used throughout the paper

E	Firm - Enterprise	e; ¬e	Enter, non-enter strategy in the PG (strategy which may be played by S_S)
S_S	Firm's strong stakeholders	PD_{Ej}	Prisoner's Dilemma(s) played in the network connecting the firm E with its weak stakeholder S_{Wj} where j = 1,2
S_{Wj}	Firm's weak stakeholder j	C_{Ej}	E's Cooperative strategy in the PDs
PG	Psychological Game involving the firm and its Strong Stakeholders	D_{Ej}	E's Non-Cooperative strategy in the PDs
U	Collusive strategy of S_S in the PG inducing an Unfair treatment	T	CSR ideal principle with which agents endowed with conformist preferences want to conform
F	The S_S's Fair strategy in the PG	λ	Exogenous parameter representing the disposition to conform with the ideal principle T
F_E	The E's Fair strategy in the PG		
U_E	The E's collusive Unfair strategy in the PG		

2.2 Modeling the network linking the firm and its stakeholders

The above analytical framework is used in this section to model the relationship of the firm with its weak and strong stakeholders. Consider the relational network of Figure 3.[9]

Figure 3 A relational Network including The Firm and its Stakeholders



A strong stakeholder (S_S), locked by mutually dependent specific investments into the transaction carried out in cooperation with the firm, and the firm ('enterprise' E) are connected by a mutual not deficient relation, while the firm E has also unilaterally deficient relations with two categories of weak stakeholders (S_{W1};S_{W2}) that, in turn, have relations with other members of the social network. To give a specific example of the network, we may imagine that: S_{W2} are employees in a plant owned by the Multinational Enterprise E in a poor developing country, where E has relocated mature productive processes for some of the items that it traditionally supplies to the global market, whereas S_{W1} is the first firm in the international supply chain furnishing components that E continues to assemble in the old plant at its headquarters located in a rich developed country. S_S may consist of high-skilled core employees at the headquarters belonging to the same local community as E's managers, well unionised and endowed with some threat power, or pension funds holding a significant share in E. Agent 3 is a second-order supplier firm (located in a developing country) within E's supply-chain (i.e. a supplier firm to E's direct supplier); agent 4 represents firm 3's employees (assumed to be better paid than S_{W2}), and agent 5 represents the developing country's retailers whose best customers are the workers belonging to 3 (whereas they

[9] This network configuration allows us to consider all the characteristics of the relationship between strong stakeholders, weak stakeholders and firms we are interested in to the aim of this paper. We will not study either other possible network configurations or the density of the relationship characterizing this network (this may be a further extension of the present analysis).

are less interested in satisfying demand by S_{W2}, who are too poor to be commercially attractive).

2.2.1. The games involving the firm and its weak stakeholders

We start the analysis of the network by focusing on the relationship between the firm and the two weak stakeholders. According to our definition of weak stakeholders, we suppose that each S_{Wj} (for j = 1,2) makes an effort to become unique to E by investing idiosyncratically in their human capital and dedicated technologies and processes, in order to increase their value to E. However, E still considers each S_{Wj} replaceable, because its main reason for relocating and having this foreign supply chain is to cut labour costs, wages, etc. Each S_{Wj} wants to maintain the cooperative relation with E, while E is not symmetrically interested in continuous cooperative relations with any of them, and seriously considers the short-term convenience of breaching at any time labour and supply chain contracts in order to relocate its plants elsewhere (where wages are even lower) or recruiting new suppliers offering components at even lower prices. Note that not cooperating does not imply for the firm the complete severing of any connection with S_{Wj}. It may merely take the form of maintaining a network of not truly cooperative relations within which E tries to expropriate opportunistically all the surplus that S_{W1} and S_{W2} may expect as the equitable remuneration of their investments. Hence, in our model, E taking all the surplus amounts to a continuing network in which E acts uncooperatively towards S_{Wj}.

To put the relation between the firm and each weak stakeholder in formal terms, we assume that they play iterated Prisoner's Dilemma Games (hereafter also PDs). The firm may cooperate or not cooperate in the PDs with weak stakeholders where:

A. cooperating means for E underwriting a long-term contract including guarantees reassuring each S_{Wj} about his/her appropriation of a reasonably equitable part of the surplus generated;

B. not cooperating means for E threatening to breach short-term supply chain contracts or incomplete labour contracts in order to extract all the surplus from S_{Wj}.

We assume that the discount rate δ_E that allows E to appreciate the long-term mutual benefits produced by S_{Wj} specific investments in term of increasing returns is

not high enough to counterbalance the short-term incentive to appropriate all the surplus, which depends on the strategic possibility of keeping salaries and prices paid to the developing country's workers and supply-chain firms very low (note that in any repeated Prisoners' Dilemma there are many possible equilibria and some of them allow substantial exploitation of one player over the other).

Finally, according to our approach, even though each S_{Wj} would like to cooperate with the firm in the PDs, they also have some defection capability (it is for this reason that we model the relationship using PDs). Weak stakeholders S_{Wj} are assumed to be able to defect (and retaliate against the firm) by using the only weapon available to them: maintaining low effort and poor quality of the goods and services provided as long as E has imperfect monitoring ability on their actions.

2.2.2 The game involving the firm and its strong stakeholder

The relationship established by E with S_S comprises various elements which, as we shall see, make a modified version of the Trust Game suitable for its formalization. Specific investments are assumed to be symmetrical and mutually dependent between the firm and strong stakeholders. E (S_S) specific investment depends for realization of its value on maintenance of the cooperative relation with S_S (E). Essentially, strong stakeholders depend for their welfare on the continuity of the cooperative relation with E; but vice versa, E depends on their cooperation for its continuing existence. This does not mean that they lack an exit strategy that interrupts or reduces the rate of cooperation, or a strategy that enables free riding on the other party's cooperative effort. In fact a key feature of the game is that S_S may choose to stay out of the interaction with E if s/he does not trust E enough to play a cooperative strategy with it. Nevertheless, continuing cooperation in this case far outweighs the discounted value of resorting to these defect strategies.

On this interpretation it is quite natural to suppose that S_S, as far as his/her material payoff is concerned, may collude with enterprise E in order to appropriate all the surplus generated by the set of specific investments made in relation to the firm. Interpretatively, we may assume that these are made by both strong and weak stakeholders, although continuous cooperation with the latter is less essential to the firm than with the former (so that expropriation of weak stakeholders may be preferred by

the firm). On the other hand, both types of stakeholder depend on the firm in order to realize their investments.

In order to capture this key point of our analysis, we model the relationship between strong stakeholders and the firm by considering a game with two active players, S_S and E, and a dummy player that ideally represents the category of weak stakeholders (S_W) affected by interaction involving the two active players. This entails that S_S and E may decide either to collude so that no resources are invested (or reserved) in order to improve the cooperation with weak stakeholders in the games that the firm will play with them in the remaining part of the network, or to treat them according to equitable terms. This means allocating part of an existing surplus for the purpose of increasing weak stakeholders' payoffs to an equitable distribution in the games that they will play with the firm in further parts of the network. We will see that the effective implementation of this decision – if it has been taken at this stage – can be interpreted as depending on a cooperative decision by the firm in the ensuing games. For the moment, however, we maintain that if this decision is taken by S_S and E, it generates payoffs also for weak stakeholders (the best interpretation is that S_{Wj} payoffs are *saved* to be given to them in the ensuing games). Here, therefore, weak stakeholders are taken as dummy players because at this stage they can only be subject to the effects of the firm and strong stakeholder's interaction, without having any voice in it. They will become active players only later, when they participate in games where they interact directly with the firm at further nodes of the relational network. Technically, this means that – with reference to the network of games in Figure 3 – the game played by E and S_S is different in form from games played by E and any S_{Wj} later in the network. Figure 4 illustrates this game in extensive form. The normal form corresponding to the extensive Trust Game is given by Figure 5.

Figure 4 The stage-game game played by the firm E and its strong stakeholder S_S - Extensive form

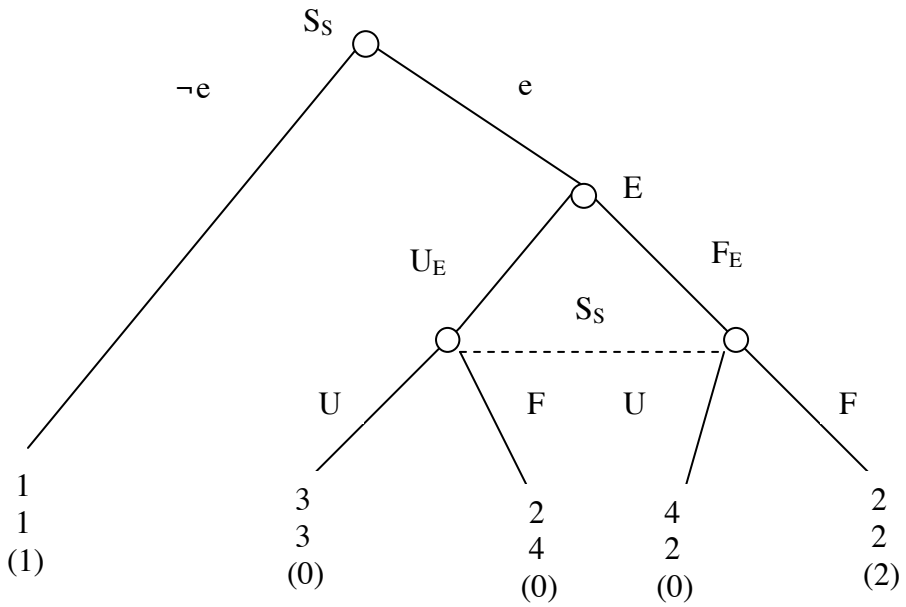


Figure 5 The stage-game played by the firm E and its strong stakeholder S_S - Normal form

E S_S	F_E	U_E
e, F	2, 2, (2)	2, 4, (0)
e, U	4, 2, (0)	3, 3, (0)
¬ e	1, 1, (1)	1, 1, (1)

In both figures, the dummy player's payoffs are reported within brackets and represent the share of a total surplus that active players refrain from appropriating so that they can pay equitable wages or prices to S_{Wj}. Thus, the dummy player's payoffs are only stakes that weak stakeholders hold in the firm's operation (payoffs are reported within brackets and the dummy player has no strategy in the game), whereas strong stakeholders not only hold stakes in the firm but also exercise influencing power.

As said, the game considered is a modified version of the Trust Game. Before S_S plays the interaction with the firm, s/he has a move where s/he may choose to enter (e) or stay out of (¬e) the relation with E. Entering means trusting E and making a specific

investment in relation to it. If S_S decides to enter into a relation with the firm, E has two possible strategies available. It may implement a collusive strategy (U_E) that allows itself and S_S to appropriate all the surplus if S_S enters and plays U as well (see payoffs (3,3,0)), or it may implement a fair division rule, F_E, that allocates a fair share to the dummy player only if S_S enters and plays F as well. This entails *saving* a share of the surplus (equal to 2) to which the weak stakeholders are entitled (see the extensive form of the game in Figure 4 and its normal form in Figure 5, where this occurs with the payoffs (2,2,2)). One-sided opportunistic behaviour against S_S occurs when S_S enters and plays 'fair' (strategy F) by restraining his/her claim, but E cheats and appropriates all the residual so that nothing is left for the dummy player. In this case we say that E is abusing S_S's trust, in so far as we understand S_S's entrance, if s/he plays (e,F), as expressing his/her intention to behave equitably toward weak stakeholders. However, one-sided opportunistic behaviour may also occur the other way round: S_S may claim the larger portion of the surplus while E moderates its pretensions. Without effective coordination on the pair of strategies F, we assume that the party which claims more by playing U is in fact able to reap the larger part of the surplus (consider payoffs (2,4,0) and (4,2,0)).

An important feature of this game is that by entering a collusive agreement (e,U;U_E), or acquiescing with the firm's opportunistic behaviour U_E, S_S puts the dummy player in a situation even worse than when S_S refuses to enter by ¬e. In other words, because of S_S's essential role in generating the firm's surplus and in allowing the firm's activity (for example, the key role of institutional investors), egoistic collusion involving both S_S and E, or at least S_S's acquiescence with E's opportunism, is strictly necessary for the complete expropriation of the dummy player. Hence a S_S that cares also for the dummy's welfare and is aware of E's devious strategy for getting around its candid self-restraint move, has an alternative for the pursuit of full fairness. This consists in boycotting E on behalf of the dummy's (second-best) stakes in the transaction. To exemplify a possible weak stakeholder situation, imagine a small firm which converts its productive plant so as to become a specialized supplier to a multinational enterprise. After the specific investment has been made, the multinational enterprise demands to change the supply contract, threatening that otherwise it will find a different supplier. This generates a situation which is worse for the supplier than the situation antecedent to the specific investment. The idea is that staying out of a relation with the firm may

prevent the strong stakeholder from inducing weak stakeholders to make specific investments that will be expropriated.

2.2.3 The S_S-E game's equilibrium solution and the instability of GM trigger strategies.

The only Nash equilibrium solution of this game is $(e,U;U_E)$, which, moreover, is in dominant strategies. This entails that the solution of this simple two-person division game is such that both players play the collusive and egoistic strategy U. Because it is the unique equilibrium point in dominant strategies of the one shot game, it will also be one equilibrium point of the repeated game that has this game as a stage-game. Hence one obvious equilibrium profile of the repeated game is for S_S (after having entered) and the firm E to adopt the iterated strategy 'play U at the first stage and thereafter, no matter what the other player does'. In the interpretive context adopted here, this solution amounts to socially irresponsible conduct by the firm with respect to weak stakeholders, while a collusive agreement is reached with the strong one (for example unions or pension funds).

The unique equilibrium in dominant strategies clarifies the extent to which this modified version of the Trust Game (TG) differs from the original TG, where the unique Nash equilibrium would be "not entering" for the stakeholder. In this case, staying out is not the S_S's best response, because "abuse" is at the expense of a third party, the weak stakeholder. In the original TG, staying out is a credible threat that the trustor may implement by means of a repeated game equilibrium strategy, if s/he believes that the trustee will play the dominant strategy of the one shot game, since it is also part of the unique Nash equilibrium of the stage game. This is not the case here, because staying out is the worst payoff to S_S, and would not be a credible move that a S_S motivated to care also about S_{Wj}'s well-being could make in order to deter adoption of a collusive unfair strategy by E.

Note that on this point our analysis significantly differs from that conducted by L&S (see Lippert and Spagnolo 2009). But it also highlights a problem inherent to the analytical framework of relational networks. L&S consider a network like the one described in Figure 3 but in which all the players' relations (including the relation between S_S and E) are modeled as iterated PDs. They state that, under perfect information and assuming that all players adopt the MG trigger strategy, a network of this kind would be sustainable (in the sense that all the players would cooperate with

each other) because of the threat of endogenous sanction against defectors implicit in the structure of MG trigger strategies. We raise a basic objection against this approach: why should player S_S implement the sanction (by stopping his/her cooperation with E) if s/he learns that the E has defected against other players in the network? Since the cooperative relation between S_S and E is mutual, and given that no other player can sanction S_S if s/he deviates from his/her MG trigger strategy, there are no endogenous material incentives for S_S to sanction E if E defects with the weak stakeholders. It seems that the MG trigger strategy would require player S_S to behave contrary to rationality, so that the sanctioning behaviour implicit in player S_S's MG trigger strategy is an ineffectual threat to player E, unable to prevent it from 'defecting' with its weak stakeholders.

The game we have introduced to model the S_S -E relationship is explicitly intended to show even more clearly the instability of the MG trigger strategy in the case of a deviation from cooperation. This problem, in fact, would entail elimination of the equilibrium based on the MG trigger strategies played by all the network participants as a sub-game imperfect equilibrium (specifically, imperfection would result, within the overall dynamic game constituted by the repeated games that any pair of adjacent agents plays in the network, from the irrationality of the behaviour required in the sub-game played by E and S_S in Figure 3).

We shall discuss this point by showing how the game specified in the previous subsection enables us to introduce a psychological game PG which in its turn will make it possible to formalize player S_S's and E's MG trigger strategy in a way that evades this instability (equilibrium imperfection) problem. This amounts to showing that cognitive social capital and the adoption of CSR principles – which we will characterize in term of the elements of the PG game – generate endogenous incentives for S_S to punish the firm if it defects against the weak stakeholders.

3. A psychological game

3.1 Conformist preferences

Our assumption is that the game played by the E and S_S described in the previous section (see Figure 4 and Figure 5) is only the basis, in terms of *game form* and *material* payoffs, for introducing the psychological game PG played by active players (the firm E and its strong stakeholder S_S) endowed with the cognitive social capital that we

associate with the concept of conformist preferences (Grimalda and Sacconi 2005 and 2007; Sacconi 2007a, Sacconi and Faillo 2010). A psychological game results directly from the former simply by adding the assumption that the players' payoffs are defined in terms of psychological utility functions (see Geanakoplos et al. 1989, Rabin 1993). Our specification of the psychological game is based on the idea of conformist preferences.

According to the conformist preferences model, agents have preferences that are defined over states of affairs described as sets of interdependent actions characterized in terms of their degree of consistency with a given abstract principle or ideal. Essentially, the model of conformist preferences is based on the idea that agents are motivated not only by material incentives, but also by the desire to conform with some ideal principle, which in the original model (proposed by Grimalda and Sacconi 2005 and 2007) is a normative principle of welfare distribution, given the players' belief in others players' conformity.

The utility function of a generic agent *i* characterized by conformist preferences is

$$V_i = U_i(\sigma) + \lambda_i F[T(\sigma)]$$

where the first term $U_i(\sigma)$ is the material utility obtained by agent *i* in state σ. The second term, $\lambda_i F[T(\sigma)]$ is the agent's ideal utility and represents conformist preferences reflecting the agent's concern for reasons to act different from the traditional consequentialist ones. Essentially, these reasons amount to a desire to conform with a normative principle *T* which is believed to be reciprocally conformed with – up to some level – by the agent itself and by the other agents participating in the same interaction through the production (by means of the agents' behaviours) of the social state of affairs σ.

First, the ideal principle *T* represents the principle on which agents agree in a pre-play communication stage under the 'veil of ignorance'. In our analysis it represents the CSR principle on which the firm and stakeholders agree from a position of impartiality and which makes explicit the firm's commitments in terms of fiduciary duties towards all the its stakeholders. In general, the formal specification of *T*, intended to express the agreed criterion of fair distribution among all the players (irrespectively off their strong or weak positions), is given by the Nash Bargaining Solution, also called the Nash Social Welfare function N:

$$T(\sigma) = N\ (U_{1,\ldots,}U_n) = \prod_{i=1}^{n}(U_i - d_i)$$

where d_i stands for the reservation utility that player i can obtain when the bargaining process collapses. Note that the status quo payoffs reflect the hypothesis that the agreement is signed under the symmetric position engendered by a 'veil of ignorance'.

Second, the weight λ_i (a positive number), is an exogenous parameter representing the maximum possible magnitude of the disposition to conform with the ideal principle *T*. The intensity of the motivation to conform with the principle *T* for agent *i* is then related to the value of λ_i. The higher λ_i is, the more agent *i* will be disposed to conform with the principle *T*, granted that it has been agreed and that agent *i* believes that the others will conform with the same principle. The parameter λ_i represents a component of cognitive social capital defined in terms of a generic disposition to conform with shared or agreed social norms, and it is taken to be an endowment of cognitive social capital (meaning disposition) that agent *i* inherits from his/her social environment (it can also be considered a biological trait fixed through evolution).

Third, the function *F* captures the effects on ideal utility of beliefs about the degree of reciprocal conformity with the ideal exhibited by the agent him/herself and other agents. *F* therefore expresses the component of our idea of cognitive social capital understood as a system of mutual beliefs on the degree of norm compliance exhibited by a given state of affairs (strategy combination) of the game. Following Grimalda and Sacconi (2005), we adopt a specification for *F* based on the hypothesis that each agent has a measure of his/her own conformity with the principle *T*, given what s/he believes about other agents, and that at the same time the agent has a measure of how much other agents' are believed to reciprocate conformity given their own beliefs.

Let us consider a two-person game. In this case, *F* can be specified by considering two elements:[10]

1. $1 + f_i$: the index of player's *i* conditional conformity. The value of f_i depends on the extent to which player *i* contributes to fulfilling the ideal *T* with his/her actions (i.e. by conforming with or deviating from the ideal), given what s/he believes about the other player's choice.

[10] See Appendix I for a formal representation of F.

2. $1+\widetilde{f}_j$: the index of player's *j* expected reciprocity in conformity, or the esteem that player *i* forms concerning *j*'s compliance with the ideal *T*. The value of $\widetilde{f}_j$ depends on the extent to which the other player contributes to fulfilling the ideal *T* with his/her actions (i.e. by conforming with or deviating from the ideal *T*), given what the second player believes (and the first player believes that the second player believes) that the first player will do.

Both f_i and $\widetilde{f}_j$ assume values from 0 to -1, so that they represent degrees of deviation from the best possible conformity with the principle *T* given the other player's (believed) action. Hence the overall utility function of agent *i* characterized by conformist preferences may be written thus (for more details see appendix I):

$$V_i(\sigma_i, b_i^1, b_i^2) = U_i(\sigma_i, b_i^1) + \lambda_i[1 + \widetilde{f}_j(b_i^1, b_i^2)][1 + f_i(\sigma_i, b_i^1)]$$

where b_i^1 is the first-order belief that player *i* has in the action of player *j*;

b_i^2 is the second-order belief about player *j*'s belief in the action adopted by player *i*.

It is clear that both conditional conformity and beliefs on reciprocal conformity as captured by the function *F*, and disposition to conform as represented by λ, play a key role in generating the (ideal) utility of player *i*. The ideal component of the utility function works as follows.

a) If *i* fully conforms with the principle *T* and believes that *j* will fully conform as well, then *i*'s ideal utility will be:

$$\lambda_i \times 1 \times 1 = \lambda_i$$

that is, the maximum possible value of ideal utility.

b) If *i* does not fully conform and believes that neither will *j* fully conform, the value of the ideal utility will be lower than λ_i:

$$(1 - x)(1 - y)\lambda_i < \lambda_i$$

c) Finally, if the conformity of at least one of the two agents is believed to be zero, then the ideal utility obtained by agent *i* goes to zero:

$$(1 - 1)(1 - y)\lambda_i = 0$$

The ideal principle *T*, mutual beliefs with regard to reciprocal conformity with the ideal principle *T*, and the disposition (λ) to conform with *T*, given such beliefs, are the components of our notion of cognitive social capital and they collapse into the value of ideal utility that the conformist agent may obtain for each give state of affairs. Hence conformist preferences equate to our definition of cognitive social capital.

As we have already noted, the disposition λ is generated by both micro and macro factors. It is connected with psychological and genetic factors that affect the disposition of each individual, and it is affected by basic social norms and cultural traits shared in the community where the agents live in a broad sense. These social norms are more general than the principle *T*, which is a principle on which agents may agree with reference to a definite domain of interactions or an organisation. Thus, while *T* is an endogenous variable determined by the players' interaction, normally engendered by their pre-play communication (agreement), λ is a contextual variable that affects the magnitude or motivational force of conformist reasons to act as they are represented by the functional *F* of the principle *T*.

3.2 A CSR principle at the basis of conformist preferences

We assume that players with conformist preferences are involved in a psychological game PG based on the modified version of the Trust Game described in Figure 4. Hence they will evaluate strategy combinations in terms of a fairness (CSR) principle *T* to which they have agreed in a pre-play communication stage of the game and whereby they make an impartial distributive justice-based assessment of the division problem that they have to solve in the game. The distributive (CSR) principle *T* is modelled as the Nash bargaining solution (NBS) of a three-person bargaining situation involving players E, S_S and a representative agent S_{Wj} – i.e. simply maximizing the product of players' payoffs net of the *status quo*. The NBS is a natural result of the assumption that E, S_S and S_{Wj} reach agreement on a distributive principle relative to the division of the surplus at stake in PG. It is not necessary that this bargaining game be taken as a game actually played. What is required is that in a pre-play communication stage the players reason 'as if' they could carry out such an agreement under the hypothesis that they cannot (or do not want to) identify with any particular player's role in the subsequent PG effectively played. Thus, in this 'counterfactual stage', they may take all the roles in

the game PG to be symmetrically interchangeable.[11] For this reason, we set the *status quo* at (0,0,0), so that all the players consider the not-fair agreement option from the point of view of the worst-off player, who would get nil if there was no impartial agreement on the surplus division. We thus express the idea that a fair agreement on the principle T *must* include all the players, and if one player gains nil from the agreement, 'behind the veil of ignorance' this amounts to not agreeing at all. Hence the two-side egoistic collusive strategy pair ($U;U_E$), or the one-side egoistic strategy U played against a fair co-operator, both signal absence of reference to any three-person equitable agreement in playing the game. This also enables the strategy ¬e to play a role in the solution, since with respect to the worst case of no distribution at all also the stay-out option with payoffs (1,1,1) could be considered a possible improvement reachable by agreement. Considering the payoff matrix reported in Figure 5, the decreasing ordering of the game states assessed according to the principle T – namely, by taking the Nash bargaining product of the payoffs corresponding to the relevant states of the game – is

1. $T(e,F;F_E) = 8$,
2. $T(\neg e;U_E) = 1$, as well as $T(\neg e;F_E) = 1$,
3. $T(e,U;U_E) = 0$, $T(e,F;U_E) = 0$, $T(e,U;F_E) = 0$.

where the last line identifies states of non-equitable agreement that are no better than the *status quo*. Note that this ordering states, as previously discussed, that S_S's staying out entails a higher level of distributive fairness in terms of Nash product than if s/he enters and acquiesces with E's collusive offer or its opportunistic endeavour to exploit S_S's fairness in order to appropriate the entire surplus.

The two active players' agreement on a principle of fair treatment including both strong and weak stakeholders amounts, in this context, to subscription by the firm to a social contract on their fair treatment – which is the core idea of CSR as we understand it. Moreover, for both the firm E and the S_S the 'fairness' strategy corresponds to a 'walk the talk' behaviour with respect to the commitment announced in the CSR norm (i.e. a code of ethics), while the 'stay out' strategy is similar to a boycotting strategy that the active strong stakeholder may (and in real life in fact does) carry out to punish companies that do not comply with the CSR commitments that they have *ex ante*

[11] Interchangeability is the obvious implication of the 'veil of ignorance' hypothesis, and allows putting aside the strategic distinction between strong and weak stakeholders and the firm.

enunciated. These intuitions are reflected by the maximum *T* value assigned to the pair of strategies (e,F;F_E), and the intermediate *T* value associated with the states where S_S decides to stay out, i.e. (¬e;U), (¬e;F).

However, it might be asked why the firm E and the strong stakeholder S_S should enter an agreement on the CSR principle *T*; and in particular what incentive E would have to do so. This question is important because – as we shall see in the next sections – in the psychological game PG that takes place after players agree on the CSR principle *T*, the player E will be induced not to abuse S_{Wj} and, consequently, to give up part of its material payoff . One could simply assume that the firm E has a value system and a corporate culture whose principles are shared by strong stakeholders and are summarized by *T*. Yet in the economic theory of the firm, 'corporate culture' is a solution for the need to acquire reputation in a context of incompleteness of contracts and unforeseen contingencies structured as a TG (Kreps 1990). In a context of this kind, the very existence of definite commitments and types functional to reputation accumulation cannot be assumed without the introduction of general and abstract principles of ethics which define, albeit with a margin of vagueness, what has to be done under unforeseen contingencies (Sacconi 2000, 2010a). In this case, the firm E must at least convince the strong stakeholders to enter the relation with it. Hence the firm must reach an agreement with strong stakeholders on general principles of fair treatment that may be employed to accumulate a reputation at least in the relation with S_S. Of course, one may say that in the one shot modified TG the firm E knows that there is a unique Nash equilibrium which entails collusion with S_S, so that E does not need any particular reputation to be able to reach such a collusive agreement with S_S. But this is not the case in the repeated game, where equilibria are necessarily multiple, and where, moreover, the commitments attached to any equilibrium strategy cannot be specified in a situation of unforeseen contingencies without recourse to general, abstract, albeit vague principles of corporate culture. Our hypothesis is that when the firm E endeavours to devise an acceptable agreement on general and abstract principles that must concern the division of a sum amongst all the three payers, the very nature of the logical exercise of formulating such principles requires it to universalize the principle of fair treatment, and hence to have exercise of the agreement cover also the weak stakeholders (which fact have no real power in the game). This amounts to saying that the agreement is reached under the veil of ignorance by active players considering the equally probable possibility of being also in the position of the weak stakeholder.

Under this hypothesis we know that the resulting agreement falls on the egalitarian solution or symmetrical NBS, as a direct consequence of impartiality when the outcome space is restricted to the equilibrium set of the repeated game (Binmore 2005, Sacconi 2010b). Besides in the theoretical literature, this result is also supported by empirical evidence on the collective choices reached by active players involved in a division problem similar to the one considered here. It has been shown that, when active players are asked to agree on a rule of division behind a veil of ignorance concerning the role that they may assume in playing the game effectively – i.e. they are faced with the possibility of occupying the dummy player's position as well – they quite directly agree on the egalitarian rule of division.

To conclude, also the assumption concerning λ may play a role, albeit an indirect one, in explaining how E or S_S can agree on the CSR principle *T*. In a context of social norms and culture wherein the presence of a high disposition to conformity (i.e. λ is high) is common knowledge, even agreeing on non-binding CSR principles with stakeholders through pre-play communication can be considered anything but 'cheap talk'. In fact, this a parameter makes it possible that conformist preferences will be formed that impinge on the players' payoff function to an extent sufficient to change the possible results of the game (of course, this could also be considered a good strategic reason for a self-interested firm not to agree at all, one to be traded off against the signal that this decision would send to stakeholders about its lack of intention to develop a reputation).

3.3. The psychological game (PG) and the ideal payoffs of players E and S_S

The previous section linked the game played by E and S_S to a basic component of the conformist preferences model: agreement on the principle *T*. However, a full description of the relevant PG game requires specification of the psychological payoffs associated with any pair of strategies. The overall utility function given in section 3.1 shows that players attach a motivational force (able to drive their practical behaviour) to something akin to 'conformity with the principle concern' – intuitively a 'deontological' motive to act – which amounts at most to a utility weight λ. This represents the maximal force of the disposition to act in conformity with the fairness principle that can counteract self-referred motives to act represented by material payoffs.

Moreover, the strength of this disposition to act in conformity with a given principle (in our case the CSR principle that implies fair behaviour by the firm towards all its stakeholders) is conditional upon beliefs that the players entertain about their reciprocal conformity with the principle. The functional *F* represents what a player deems to be the overall degree of conformity as based on the combination of the two personal indexes of conformity attached to players' decisions in relation to the principle. Taking S_S's perspective, these indexes state:

(i) the extent to which S_S conceives him/herself to be conforming by choosing any particular strategy, given his/her belief about E's strategy choice, and

(ii) the extent to which S_S thinks player E conforms by means of any particular strategy that s/he believes E may choose, given S_S's second-order beliefs about E's belief in S_S's choice.

Recall that the values of the two conformity indexes result from the subtraction of a deviation measure ranging between 0 (no deviation at all from the principle) and -1 (complete deviation) from the unit (i.e. 1 means maximal conformity), and consider in turn the different possible belief systems (i.e. first- and second-order beliefs) justifying the prediction of any given outcome of the game. Then the conformity indexes attached to how players carry out each state of the game (and consequently their ideal utility) may be computed with reference to the basic game form given in Figure 4 and 5, keeping track of the *T* values computed for each strategy combination given in section 3.2. [12]

Let us start by considering the ideal utility to be added to the material payoff of player S_S because of his/her conditional conformity index and the expected reciprocal conformity index of the firm, namely $1 + f_{Ss}(\sigma_{Ss}, b^1_{Ss})$ and $1 + \tilde{f}_E\left(b^1_{Ss}, b^2_{Ss}\right)$, as they are specified at each possible state of the game. Consider first the strategy σ_{Ss} = (e,F) of player S_S given his/her first-order belief that E plays F, (b^1_{Ss} = F_E), and his/her second-order belief that E believes that S_S plays (e,F), (b^2_{Ss} = (e,F)). The index of conditional deviation of player S_S is

$$\frac{T(e,F;F_E) - T^{MAX}(F_E)}{T^{MAX}(F_E) - T^{MIN}(F_E)} = \frac{T(e,F;F_E) - T(e,F;F_E)}{T(e,F;F_E) - T(e,U;F_E)} = 0,$$

In fact, given that E plays 'fair' F_E, for S_S by responding with (e,F) the best *T* value is attainable, which entails a conditional conformity index $1 + f_{Ss}(e,F;F_E) = 1$. For the same strategy pair, by symmetrical reasons, the expected reciprocal deviation of player E is

$$\frac{T(F_E;e,F) - T^{MAX}(e,F)}{T^{MAX}(e,F) - T^{MIN}(e,F)} = \frac{T(F_E;e,F) - T(F_E;e,F)}{T(F_E;e,F) - T(U_E;e,F)} = 0,$$

which entails that the expected reciprocal conformity index of player E is $1 + f_E(F_E;e,F) = 1$. Thus the ideal utility of player S_S for this strategy combination is the full weight λ (namely, $1 \times 1 \times \lambda$).

By the same method, S_S's conditional conformity indexes and E's expected reciprocal conformity indexes can be computed for each strategy pair, and the ideal utility of player S_S can be derived (see the appendix to this chapter for calculations). The results are the following:

- Player S_S's strategy (e,F), given his/her first-order belief that E will play U_E and his/her second-order belief that E believes that s/he will play (e,F), obtains ideal utility 0 for S_S. In fact, against a player E who unfairly plays U_E, entering and playing "fair" by (e,F) gives the worst *T* value, which is equal to 0 with respect the best possible alternative of "staying out" by ¬e, which gives a *T* value equal to 1. Recall that a single conformity index equal to 0 entails that ideal utility is nil.
- Player S_S's strategy (e,U), given his/her first-order belief that E will play F_E, and his/her second-order belief that E believes that s/he will play (e,U), obtains ideal utility 0 for S_S. In fact, against a player E who plays "fair" by F_E, responding with (e,U) means selecting the worst *T* value, which is equal to 0, with respect to the better alternative of responding fairly by (e,F), with *T* value 8.
- Player S_S's strategy (e,U), given his/her first-order belief that E will play U_E and his/her second-order belief that E believes s/he will play (e,U), gives ideal utility 0 to S_S. In fact this choice entails "collusion" with the worst *T* value, equal to 0,

[12] See Appendix I for a complete application of the calculation method.

whereas responding by "staying out" would give a better *T* of value 1, which is also the best given player E's choice.

- Player S_S's strategy (¬e), given his/her first-order belief that E will play U_E and his/her second-order belief that E believes that s/he will play (¬e), gives S_S ideal utility λ. In fact, responding by (¬e) to E who plays U maximizes the *T* value, so that player S_S's deviation is 0. At the same time, given that player S_S "stays out", player E cannot do any better in order to maximize *T* than choose one or other (indifferently) of its two strategies U_E or F_E, since both of which give a *T* value equal to 1, and both of which have a deviation index 0. However, if E chooses F_E, choosing (¬e) would no longer induce a conformity index 1, because S_S in this case could maximize *T* by choosing (e,F).
- Player S_S's strategy (¬e), given his/her first-order belief that E will play F_E and his/her second-order belief that E believes that s/he will play (¬e), obtains ideal utility $1/8\lambda$. In this case, given that E plays "fairly" by F_E, player S_S does not maximize the *T* value by "staying out". However, nor does s/he minimize it, since the worst *T* value equal to 0 would be reached if s/he played unfairly (e,U). Player S_S thus scores a high deviation index -7/8, and hence his/her complementary conditional conformity index is low, that is, 1/8. On the other hand, player E, who believes that player S_S stays out, cannot do any better in order to enhance the *T* value than playing one or other (indifferently) of its strategies, F_E or U_E. Thus by playing F it obtain its maximum *T* value conditional on the (¬e) choice by S_S. So the E's expected reciprocal conformity index is 1, which combined with 1/8 allows only an ideal utility $1/8\lambda$ to enter player S_S's overall payoff for this state.

To sum up, the only way for S_S to be fully conformist is to 'enter' and choose 'fair' if s/he predicts that also E plays 'fair', but to stay out otherwise. This latter behaviour is an important consequence of the conformist preference model: staying out of an unfair cooperative relation can induce the relative best level of conformity if the other player's 'cooperative' choice is such that acceding to such a proposal of unfair cooperation or collusion would induce a lower level of implementation of the principle *T*. Thus accepting whatever level of cooperation or collusion, if it is unfair in terms of the principle *T*, is not supported by conformist preferences. On the contrary, a "principled" refusal to interact can be supported by conformist preferences, which translates into an

endogenous psychological incentive to punish the other party's unfair choices. On the other hand, by 'staying out' when E chooses 'fair', the strong stakeholder S_S permits only poor implementation of the principle. Finally, compliance would be nil not only if S_S colludes, but also if s/he acquiesces with E's opportunism by candidly choosing 'fair' when E is getting around its 'pure' intention by playing U_E to appropriate the entire surplus.

Thus far, things have been considered from S_S's perspective. Note, however, that player E's index of conformity and its index of expected reciprocal conformity about player S_S are derived by combining the same strategies described above. For example, E's index of conditional conformity $1+f_E(\sigma_E,b_E^1)$ is based on the identical strategic combinations taken into account by player E's expected reciprocal conformity index $1+\widetilde{f}_E\left(b_{Ss}^1,b_{Ss}^2\right)$ as seen in the eyes of player S_S – since the first-order beliefs of player S_S consist of player E's strategies, and his/her second-order beliefs about player E's beliefs equal player E's beliefs about player S_S's strategies. Then the two indexes must have the same values. Situations and payoffs, considered according to player E's or player S_S's beliefs of first and second order predicting such combinations are perfectly symmetrical for the strategies pairs (e,F;F_E), (e,F;U_E), (e,U;F_E) and (e,U;U_E). Of course, player E does not have move e, but it is ineffectual with respect to the symmetry of the situation that occurs after player S_S's 'entrance'. Then E's indexes of conditional conformity and expected reciprocal conformity must be respectively identical to those just considered for S_S; hence also the ideal payoffs must be the same. The only situations left to consider are those that cannot be symmetrical between players E and S_S, namely (¬e,F), (¬e,U) – i.e. situations where E's first-order belief predicts that player S_S will choose (¬e) while E's second-order belief is that S_S believes that it will choose either F or U. In these cases

- Player E's strategy F_E, given his/her first-order belief that S_S will play ¬e and its second-order beliefs that S_S believes that E will play F_E, obtains ideal utility equal to $1/8\,\lambda$. In fact, when E predicts that S_S will stay out, it cannot do any better to maximize the *T* value than choose whichever of its strategies F_E or U_E. However, what reduces overall conformity in this case is the expected reciprocal conformity of S_S, which is at the poor level of 1/8 (consider that his/her best conformity index would be associated with playing (e,F), while the worst one would be given by

playing (e,U)). The result is $1\times1/8\times\lambda$, which is the ideal utility that enters E's payoff for this outcome.

- Player E's strategy U_E, given his/her the first-order belief that S_S will play ¬e and his/her second-order belief that S_S believes that E will choose U_E, obtains the highest ideal utility λ. In fact, also in this case player E is doing as much as possible to maximize the *T* value, given the ¬e choice by S_S (since by staying out S_S frustrates any attempt by E to affect the result). But in this case this also applies to S_S, who predicts that E will in fact choose U_E, and hence rightly chooses to stay out, which makes the *T* value equal to 1, whereas if s/he had 'entered', that value would have been only 0 (in the case of both bilateral or unilateral collusion) .

To be noted in regard to these last two points is that, symmetrically with what we said concerning the motivational force of S_S's decision to 'stay out', when S_S is commonly predicted to play ¬e, the individual responsibility of player E concerning the level of principle attainment is nullified. E cannot do anything about the level of *T*, which cannot deviate from the one determined by player S_S's decision. Since E cannot be responsible for any deviation from the level of *T*, conformity is intact and maximal whatever the choice of E (U_E included). This may also be understood in the sense that, by staying out, S_S prevents any deviation from conformity that might be attributable to a choice by player E, whose intentions cannot be relevant in terms of responsibility, as far as they are at all fancyfull (E knows that, whatever its virtual choice, the game is over due to ¬e) and ineffective with respect to the game's outcome. In any modified TG, such as the one under consideration, ¬e entails that the game ends before E's decision node has even been reached. However, the conformity index is not a measure of a player's counter-factual intentions, but only a measure of the factual deviation due to his/her decision from the best reachable level in terms of a given standard, conditional on the other players' behaviour. It takes the dictum "ought implies can" quite seriously, and in this case player E cannot be considered responsible for any deviation from the given level of conformity with the principle *T* set by player S_S. A different conclusion would be admissible if E assigned a positive probability to S_S not being truly playing ¬e. But this hypothesis is not admitted under the psychological games assumption that beliefs are internally consistent with their psychological equilibria and are common knowledge among the players. Hence it is admissible for conformity indexes in these

cases to assign a zero deviation to any choice by player E and hence full conformity with player E's choices. However, the case of player S_S is quite different. When player S_S predicts that player E will choose U_E at his/her decision node either because its plans to collude or because its already knows that S_S will stay out and hence feels relieved of any decisional responsibility toward *T*, then s/he is fully responsible for prevention of the possible effect of the predicted decision by E on *T* attainment. Hence, in order to conform with the principle, s/he must play ¬e. This is reflected in the best S_S conditional conformity index (or in the best expected reciprocal conformity index, as seen in the eyes of player E), which is equal to 1 for that choice by player S_S.

To sum up, in correspondence to each combination of strategies (states of the game) conditioned on a system of consistent first- and second-order beliefs (i.e. beliefs predicting exactly the state of the game under consideration), for every player we can single out the values of the conformist component of his/her utility function by computing the relevant combination of both the conformity indexes of a player.

Before continuing with discussion of the psychological equilibria resulting from integration of material payoffs with ideal utilities deriving from conformists preferences, we give some intuitive substance as to why we consider the possibility that a firm may have a positive psychological payoff from applying an ethical principle of cooperation with all its stakeholders. Here our approach is closely linked with Aoki's notion of corporate social capital (Aoki 2010, *infra*): 'Corporate social capital may not be immediately cashed in, but it may be enjoyed by various corporate stakeholders in non-pecuniary manner, e.g., the pride of employees working for a socially reputable corporation, satisfactions of environmentally-conscious stockholders from owning 'green' stocks, amenities of citizens living in clean local community and the like.' λ_E may be interpreted as the psychological payoff obtained by those with residual control rights (the owner or the top management in case of public companies), who may have conformist preferences and may obtain a positive psychological payoff from adopting corporate responsible behaviour.

3.4 Psychological Equilibria in the PG

Given the different values of ideal utility deriving from conformist preferences, the normal form of the psychological game with conformist preferences is shown in Figure 6.

Figure 6 Normal form of the PG game played by S_S and E

E / S_S	F_E	U_E
e, F	$2+\lambda_{Ss}, 2+\lambda_E, (2)$	2, 4, (0)
e, U	4, 2, (0)	3, 3, (0)
¬ e	$1+1/8\,\lambda_{Ss}, 1+1/8\,\lambda_E, (1)$	$1+\lambda_{Ss}, 1+\lambda_E, (1)$

The generalized form of this game under the assumption that payoffs satisfy the conditions d >c > b > a, is depicted in Figure 7.

Figure 7 Normal form of the PG game played by S_S and E – generalized form

E / S_S	F_E	U_E
e,F	$b+\lambda_{Ss}, b+\lambda_E, (b)$	b, d, (0)
e, U	d, b, (0)	c, c, (0)
¬ e	$a+k\lambda_{Ss}, a+k\lambda_E, (a)$	$a+\lambda_{Ss}, a+\lambda_E, (a)$

Where $0 \leq k \leq 1$

It is evident from inspection of the psychological payoffs that, in general, if λ_E and λ_{Ss} are both >d – b and λ_{Ss}>c – a (with the particular specification of payoffs parameters with which we have worked thus far, however, both conditions collapse to

λ>2), there are three Nash psychological equilibria under conformist preferences: (e,U;U_E), (e,F;F_E) and (¬e;U_E). Most interesting are the equilibrium strategy profiles (e,F;F_E) and (¬e;U_E). Each of these must be understood as being contingent on the respectively appropriate system of mutually consistent beliefs of first and higher order. In regard to the former, player S_S must be believed to be playing (e,F) and player E must be believed to be playing F_E, while both of them must believe that the other has exactly these beliefs (and the consistent beliefs about beliefs). When these conditions are satisfied, the conformist payoffs reported in the upper left cell of the normal form game in Figure 6 are effective (because they depend on indexes of conformity contingent on exactly these beliefs) so that if λ_E and λ_{Ss} are both >d – b, the players' mutual best responses are (e,F) and F. This means that both players have a desire to conform with their ideal principle of justice sufficient for them to prefer forgoing a material self-interested benefit achievable through a collusive agreement in order to ensure fair treatment of the dummy player.

Because of the existence of the second equilibrium, S_S must be believed to stay out and E must be believed to play U_E, while both of them must believe that these beliefs are also held by the counterparty and that they know what the other believes. When these beliefs are satisfied, the psychological conformist payoffs reported in the bottom right cell of Figure 6 are effective, so that ¬e is S_S's best response to E's strategy U_E (which in turn is its best response to ¬e). Note that the condition for the existence of this second equilibrium is λ_{Ss}>c – a , which is not required for λ_E. This is required only of player S_S since the decision to stay out of the cooperative relation with E is his or hers alone. Intuitively, this implies that, by trading-off conformist utility with material payoffs, S_S prefers to boycott E more than collude with him/her. Essentially, the 'sanction strategy' of strong stakeholders, which have a key role in inducing the firm to be 'fair' with weak stakeholders by respecting the CSR principles, does not require any condition on the firm's psychological payoffs which may be 0. Consequently, even firms which are not intrinsically motivated by cognitive social capital, if they agree on CSR principles in order to induce their stakeholders to undertake optimal investments, may suffer be sanctioned by strong stakeholders (if the value of λ_{Ss} is high enough) and may be induced to cooperate with weak stakeholders.

Finally, a further psychological equilibrium is the old Nash equilibrium $(e,U;U_E)$, which materialises when the previous conditions on beliefs systems are not fulfilled even if the conditions on λ_E and λ_{Ss} are satisfied. That is, notwithstanding the absolute potential of the disposition to act in accordance with the principle of justice, this equilibrium emerges when mutual confidence in reciprocal effective conformity breaks down. This amounts to a beliefs system such that player E neither believes that S_S is effectively compliant with the principle, so that s/he would really play strategy (e,F) if E were to choose strategy F_E, nor is confident that S_S would really play strategy $\neg e$ if it offered collusion by strategy U_E. At the same time, S_S neither believes that E will play F_E when s/he plays (e,F) nor believes that player E is confident that s/he will really play $\neg e$ if E plays U_E. Under these conditions of mistrust, an S_S playing $\neg e$ would act against the systems of mutually consistent beliefs that predict $(e,U;U_E)$ as the result, which is not admissible in terms of psychological equilibrium. In the absence of beliefs systems that justify playing one of the other two psychological equilibria, $(e, U;U_E)$ emerges as the only psychological equilibrium, even though it is based on just material payoff**.**

4. Cognitive social capital and the endogenous sustainability of cooperative networks of relations

The psychological game PG played by E and S_S reveals the importance of both cognitive social capital and of CSR principles in allowing the endogenous sustainability of cooperative relations between the firm and all its stakeholders that were considered as mere possibilities – far from being effective – in networks like the one reported in Figure 3. In this section we set out the main result. A rigorous proof must wait for the next sections.

4.1 Cognitive social capital as conformity disposition

A high level of cognitive social capital in terms of disposition (λ) is a necessary, even if not sufficient, condition for obtaining structural social capital between the firm and all the stakeholders. If the conditions on the parameter λ are not satisfied, only the unfair or collusive equilibrium $(e,U;U_E)$ can emerge. Referring to the distinction between bridging and bonding social capital - 'There may be high social capital within a group ('bonding' social capital) which helps members, but they may be excluded from other groups (they lack 'bridging' social capital.' (Narayan 1999, p.3) - we may say that

the collusive equilibrium is an example of bonding social capital (between the firm and the strong stakeholders), while the fair cooperative equilibrium between the firm and its strong stakeholders is an example of bridging social capital. Sufficiency conditions for bridging social capital include both dispositions and beliefs systems. Bonding social capital obtains whenever the disposition to conform with impartial norms is insufficiently strong or when, owing to contingent conditions, expectations of mutual distrust emerge concerning reciprocity in conforming with fair and impartial norms, whereas players have consistent beliefs systems that allow them to predict collusion (which, moreover, must be a Nash equilibrium). This characterisation of bridging social capital in terms of equilibrium conditions shows that, even though some of its components may be objectively determined at the level of the biological or cultural heritage of a given category of individuals, most of it is nevertheless relative and contingent on fragile conditions of social interaction amongst rational individuals. Beliefs systems in particular exhibit this contingency, for there is no absolute reason for some of them to be completely discarded so as to ensure that only the 'desired' beliefs systems emerge to support good equilibria. In fact, how could we exclude *a priori* that a situation of mistrust may emerge even amongst people with the highest disposition to conform with social norms and ethical principles?

4.2 What affects beliefs in the PG. The role of agreement

As usual, multiple equilibria (especially multiple psychological equilibria) make any prediction about the effective solution of the game depend on the availability of an equilibrium selection mechanism able to explain the formation of any given system of mutually consistent beliefs whereon equilibria are contingent. This is not a matter of brutally biological or traditionally determined cultural inheritance. On the contrary, equilibrium selection depends on fragile cognitive mechanisms of belief formation, such as how individuals reasonably react to different choice contexts and how they learn from past interactions. Far from being able to uniquely answer this problem, conformist preference theory is not completely mute about it. Recourse to the 'cognitive role' of ethical norms and distributive justice principles helps give partial predictability to the emergence of the system of beliefs required for bridging social capital to be created and the corresponding fairness equilibrium to be implemented (see Sacconi 2010c, *infra*).

Modelling the game in terms of conformist preferences entails some implicit assumptions. In particular, as already said, it amounts to assuming that, before this game is played, there must be a phase of pre-play communication (traditional game theory would rule it out as 'cheap talk', but we shall see that it is quite important in affecting the players' preferences). In this phase, players adopt the cognitive perspective of an ideal game 'under a veil of ignorance' such that they are able to agree impartially on a norm or a principle of fairness which they deem relevant to the distribution of surpluses generated in interactions like the one involving E and its stakeholders. The 'impartiality' of this point of view consists in the fact that, with ignorance of who will take *ex post* whatever role in the game (be it the role of E or the role of whichever category of stakeholders), a (CSR) principle of fair division is *ex ante* agreed upon by anonymous players in order to establish how the real life division game will be played *ex post*. This may be seen as reasoning 'as if' the players were involved in a fictitious bargaining game 'under the veil of ignorance'. But alternatively it may also be seen as simply a cognitive process of reasoning whereby players are detached from the personal perspective and their interests in the concrete situation, and simply recognize that the situation (the game) they are going to play has to be categorized as one element pertaining to a more general class of situations where a given principle or social norm of fairness is normally applied. Put differently, the situation exhibits to a significant degree the pattern or the silhouette of a category –or fuzzy membership of a set – which is normally understood as the domain of application of a given principle or norm of justice.[13]

What is distinctive of this pre-play communication stage is that in one way or another it operates as a *framing effect* on both players' motivations and beliefs. According to the motivational point of view, framing a situation as one involving a fair agreement on a principle of justice activates a motivational drive (what we may call a disposition to act in conformity with a mutually agreed principle) able to produce a specific behaviour or the 'desire' to be just. The intensity of this 'desire', or the causal

[13] These are just two different ways to approach the same point, however. In fact players could not categorize the situation as one whereon an impartial principle of justice normally applies if in some sense they would not envision it as if they were 'under a veil of ignorance'. A situation wherein an individual performs a format of reasoning such that independently of the consideration of his/her individual identity s/he is capable of agreeing on a principle of equitable distribution with other individuals supposedly similarly detached from the urgency of their material claims, is quite similar to the cognitive process whereby s/he performs the task of subsuming the concrete distributive case under a more general and

force of this disposition seen as a preference (which is a sort of *passe-partout* for intending whatever motive to act) is what the model captures with the parameter λ. Hence, it is because in a pre-play communication phase the situation has been assessed in terms of an impartial agreement or according to a commonly shared principle that in the "real life" game players may frame the situation so that they feel the motivational force to act in accordance with it "up to level" λ.

From the cognitive point of view, framing the situation as one of impartial agreement, or simply as an exemplar of a wider category to which a general abstract principle of impartial treatment applies, affects the players' beliefs. When a situation is recognized as belonging to an abstract category requiring impartial treatment, the individual reasoner proceeds *by default* to the position that there is no reason or evidence for not believing that both him/herself and the counterparts will envisage the situation in the same way. The abstract norm or principle (in our case the agreed CSR principle) defines a mental model of *the* rational agent as a typical agent that agrees on a principle and hence is (until proof to the contrary) committed to it, or as an agent who behaves as normally observed within a category of cases subsumed within the domain of a norm or principle. 'People that voluntarily agree on a principle or who understand this situation as belonging to a category identified by the validity of a norm, normally behave like that…' – this defines a normative mental model of agent that the individual reasoner endorses under the framing effect of what we called the pre-play communication phase (see Sacconi and Faillo 2010).

There is no definitive proof that all agents will actually act according to this model. Rather, it is the simplest model of agent that follows from the fact that the situation has been framed as a situation of impartial agreement or a case belonging to a general class identified by a norm of justice. It might be said that if one freely agrees to a principle, one expresses the plan or the intention of acting according to the provisos of the agreement itself. Hence, until proof to the contrary, one may expect the rational agent to act 'normally' according to his/her free agreement. If one categorizes a situation as a case in a class subsumed under the domain of an abstract norm, the norm defines how people *normally* act within the category (or must act to stay in it) until proof to the contrary. Hence, one has a mental model of how people normally behave (or normally

---

abstract principle of justice such that the case will be treated according to the impartiality criteria inherent to the principle.

should behave to satisfy the premise of an impartial agreement or consistency with the normative statement of a norm) under the current categorization, until proof to the contrary.

Admittedly, all these are just *default* inferences, valid under *caveats* such as 'normally', 'until proof to the contrary' , 'not contrary to what we already know' etc. But they are nevertheless perfectly legitimate within their limits. If these are the stereotypes of a rational agent under the current framing of the situation, they are also the mental models that 'come to the agent's mind' when s/he tries to decide rationally, those that s/he takes for granted or as provisionally valid to plan his/her action. There is no conclusive reason for doing this except that these constitute the model of the rational agent that comes to his/her mind under the current framing effect.

Now imagine that the same agent is asked to forecast the behaviour of other agents (for example the second player in the real life game). In the absence of contradictory information or evidence to the contrary, by default s/he will simulate the other agents' reasoning and behaviour by applying the same mental model used to provisionally define his/her own plan or conduct. The rational basis for this replication has the same fragile but nonetheless intelligible basis as before: the simplest way to forecast other agents' behaviour, as long as there is no evidence or proof to the contrary, is to deduce their behaviour from the best mental model of an agent inferred from the frame of the situation. '*Assuming that the situation has been understood as one of impartial and generally acceptable agreement, or one normally categorised as the domain of application of a neutral norm, given that I need to work out a forecast of other agents, I do not find any reason not to apply to these other agents the same mental model that seems valid for myself as it is consistent with a norm which is independent of any characteristics that make me different from any other*'. As long as there is no evidence that other players do not participate in the same impartial agreement or do not categorize the current situation in like manner, by default we conclude that the same model of agent that came to our mind to define our action is also valid for symmetrically forecasting other agent's decisions and behaviours.

Given the mental model just described, if players participate in the pre-play communication stage (the agreement on CSR principles) their first-order beliefs in the psychological game consist of the mutual prediction that strategy choices are (e,F) and (F), and their mutual second-order beliefs are hence consistent with these predictions about choices.

4.3 Cognitive social capital and 'modified' MG trigger strategies

The analysis of belief formation resulting from the pre-play communication phase provides a sound and workable starting point for our model, but no more than that. In fact it works only in a one-shot game, where there is no previous experience and no evidence can be uncovered that contradicts the mental model derived from the ideal choice or the abstraction and categorisation process carried out at the pre-play communication stage. However, when the game is repeatedly played, observations of previous effective behaviours necessarily influence beliefs about what strategy the counterpart is effectively playing.

Here we make our first basic assumption about the dependence of player S_S's beliefs and behaviours in our psychological game (the PG) on what s/he learns from the behaviours of player E in the other games in which it participates through the relational network considered in Figure 3. We call all of them PD_{Ej} in order to indicate that they are Prisoner's Dilemmas played by E in relation to player $S_{Wj} = j$. We assume that

A1) If S_S learns that player E defects at time t in a PD_{Ej}, s/he understands that E is not 'really' playing the strategy F in the PG from that stage onwards.

In fact what has been *saved* and entitled to S_{Wj} in the solution of the component game of division of the surplus PG has not been used to remunerate players S_{Wj} equitably by cooperating with them. Thus, at stage $t+1$, S_S will predict that player E is not playing 'fair' in the current repetition of PG. This signifies that the condition for the emergence of the 'no entry' psychological equilibrium has been activated (obviously, the 'no entry' decision depends also on the value of λ). Of course, this point is particularly important in relation to the strategy ¬e that is what in our model takes the place of the punishment stage strategy discussed in section 3.4, and the psychological equilibrium involving ¬e seems to be what we needed to show that implementing the punishment phase in player S_S's strategy is compatible with S_S's (conformist) incentives. To guarantee this result, however, we need not only to show that, when s/he learns about a defection against weak stakeholders, S_S believes that E will choose U_E in PG at $t+1$. We also need to show that E predicts that S_S will not enter at time $t+1$. The 'no entry' equilibrium is contingent on this reciprocal beliefs system.

Assumption *A1* requires a *caveat*: S_S does not understand that E is not really playing the Fair strategy F_E in the PG when it defects for the first time in the PDs with weak

stakeholders, in case this is required by implementation of E's MG trigger strategy. It seems in fact likely that E does not lose his/her trustworthiness in the eyes of S_S if s/he knows that E is required to defect by compliance with a MG trigger strategy itself intended to support cooperation throughout the network. Nevertheless, this forgiveness cannot last for more than one period because also player S_S's sanctions based on how s/he assesses player E's behaviour are needed in order to provide player E with the appropriate incentive not to take advantage of its relation with S_S to exploit weak stakeholders.

To suppose that E realizes that S_S will not enter at *t+1* after it has defected against weak stakeholders is quite intuitive in a context where players have first agreed on a fairness norm and have also conformed with it, so that player E knows that S_S has effective conformist preferences encapsulating a desire to be consistent with a shared norm of fairness. However, having a strong disposition to conformity is not enough if the relevant beliefs do not exist as well. This hypothesis must be rigorously justified, for the emergence of a psychological equilibrium in a given stage game depends strictly on the players' reciprocally consistent beliefs. Here we introduce our second assumption concerning the link between the equilibria of the game PG and how other games are played by different players throughout the network. We assume that S_S plays each PG stage game by following a version of the multilateral grim (=MG) trigger strategy.

A2) S_S at first plays (e,F), but after some stage t s/he plays ¬e if s/he learns from a defection occurring at stage t-1 in a PD_{Ej} – which E plays with any S_{Wj} – that E is not going to play Fair in the current PG (under the same caveat valid for assumption A1).

The strategy adopted by S_S to play his/her repeated game as a function of E's past behaviour is common knowledge in the network. This entails that once, at whatever stage *t* in a repeated game PD_{Ej}, player E chooses to defect, it also obtains the information that player S_S will play ¬e in the following stage *t+1* of the PG. But this is exactly the basis for the E's belief that at *t+1* S_S will play ¬e, and for S_S's second-order beliefs that E predicts that s/he will stay out at stage *t+1* – i.e. the condition for emergence of the 'no entry' psychological equilibrium at *t+1*.

The *caveat* to *A1* is again relevant. Also E's strategy is common knowledge, so that S_S knows whether E will adopt a MG trigger strategy such that if at *t-1* a defection occurs in the network, then player E will play 'defection' in the PD_{Ej} at the stage *t*. But

it is *not* required by assumption *A1* that player S_S immediately anticipates that player E is not going to play consistently with the fair strategy at stage *t*. This understanding can be delayed until after E's defection effectively occurs, so that player S_S, given his/her state of information and repeated strategy, must start to play ¬e at stage *t+1*. Player S_S believes that E predicts that s/he will change her choice at *t+1* and also that s/he realizes that E defected at *t*. At the same time, E predicts that S_S will change his/her strategy at *t+1* and also believes that s/he realizes that E's change of strategy occurred at *t*. Mutually consistency of beliefs is satisfied in order to allow the emergence of the 'no entry' equilibrium profile at *t+1*.

As a consequence, we are not assuming that S_S should implement the MG trigger strategy as a rule follower without having the proper psychological incentive to do so (as noted in sub-sec. 2.2.3). On the contrary, the sanctioning strategy adopted at the *t+1* stage in game PG has a perfectly endogenous explanation. The adoption of the multilateral grim trigger strategy is perfectly consistent with the equilibrium behaviour that S_S implements in the stage-game in which the strategy requires him/her to sanction E. We may say that player S_S plays ¬e *because* s/he is believed to follow the multilateral grim trigger strategy as a function of E's behaviour, but the content of this belief is now perfectly consistent with the psychological equilibrium behaviour that s/he implements in the game.

In conclusion, although we have not still precisely worked out the relation between what happens in a single PG stage-game and the strategies played in the repeated games that take place throughout the network, we have laid the bases for answering the central question: why should S_S carry out his/her threat to punish E if the latter had failed to cooperate with some S_{Wj}? Our answer is that, under the proper beliefs about S_S, s/he is ready to act as a conformist agent also if E continues not to conform with the agreed norms. Hence punishing player E by 'staying out' in the current stage game PG, is perfectly in line with player S_S's psychological incentive (when λ is high enough to counterbalance the material payoff). By anticipating S_S's behaviour, given our assumption (on belief formation and value of λ[14]), the firm E will also have the

[14] When we move from the one-shot game to the iterated interactions between the firm and its stakeholders, the possibility that λ could endogenously change with the games' result may be taken into account. It could be assumed, for example, that λ of S_S and E increases at each stage when they experience conformity indices equal to 1 or 1-ε. Our analysis does not consider this possibility, which could represent an extension of our model.

incentive to avoid opportunistic behaviour against weak stakeholders so as to prevent S_S's retaliation.

This suggests (even if a rigorous proof must wait until the next two sections) that cognitive social capital, as understood here in terms of conformist preferences and the related systems of consistent beliefs, is at the very root of the possibility to make cooperation sustainable in a relational network of repeated games, which is what we typically mean by the term 'structural social capital' seen as a set of effective cooperative relations based on trust.

5. Strategies and beliefs formation in the psychological game as a function of repeated playing of games in the relational network

The aim of this section is to provide a clear link between the one-shot psychological game (PG) played by E (the firm) and S_S, discussed in section 3, and the framework of network analysis reported at the beginning of section 2. Hence, here we consider the PG as a stage-game within the repeated playing of the games (not only repeated PG but also other games) in which players are involved throughout the network. Our aim is to adapt the MG trigger strategy to the roles performed by E and S_S in the repeated playing of games in the network: that is, its specification in consideration of the peculiar game in which players S_S and E are involved – the repeated PG .We use the analytical framework introduced in section 2.1 and we refer to the notion of sustainability of a relational non-mutual network as set out in L&S (2009). We introduce a variation of the MG trigger strategy that will account for how this strategy is specified with reference to the manner in which the repeated PG must be played in function of behaviours maintained in games nested in each other throughout the network so that it can support cooperation in all these repeated games.

To this end, we first need to identify the strategy profile of a player *i* involved in the network described in Figure 3, which comprises players E, S_S, S_{W1}, S_{W2} and agents 3, 4 and 5, that at each stage participate in playing the repeated games (normally, but for S_S and E, two adjacent games) in the network. It should be borne in mind that S_S plays only an iterated PG with E, while all the other agents play two iterated Prisoners' Dilemmas with adjacent agents belonging to the network. As a consequence, only player E is involved in three games (the PG and two PDs) at each stage.

We define h^t as a history of all the repeated games played by the agents belonging to the network. h^t is one of the possible sequences of moves available to players until the

period t. The set of all the possible histories h^t is termed H^t. Player *i*'s strategy is defined as a function that, at any time *t*, associates with each history $h^t \in H^t$ the moves that will be selected by player *i* from *t+1* onwards: $s_i : f(H^t) \rightarrow A_i^{t+1} \ \forall t$.

Note that the strategies of an agent *i* who plays a repeated Prisoner's Dilemma in our network do not only depend on the decisions made by *i* and the players who play the game with him/her. They are also determined by the moves made by the other agents in the network, even though they are not directly connected with *i*. In fact, the MG trigger strategy, which we assume to characterize the way in which these games are played, implies that every player $i \in N$ starts cooperating with his/her neighbours, and continues to cooperate as long as s/he observes that all the other players cooperate. But s/he stops cooperating if s/he observes that someone, somewhere in the network, defects. Moreover, the strategies of the firm E and of the S_S also depend on the history that characterizes the psychological game in which they are involved and which is different from the PDs played in the rest of the network. This amounts to saying that both the enterprise's and the strong stakeholder's strategies in the psychological game are a function of the Cartesian product of the histories which come about both in the psychological game and in all the repeated prisoner's dilemmas:

$$s_i : f(H_{PG}^t \times H_{PD1}^t \times H_{PD2}^t \times ... \times H_{PDn}^t) \rightarrow A_i^{t+1} \ \forall t \qquad (i = \mathrm{E}, \mathrm{S_S})$$

where: H_{PD1}^t is the set of all the possible histories which may hypothetically characterize the PD_1 - i.e. the repeated Prisoner's Dilemma between the firm and the first agent connected with it in the network. In regard to the network depicted in Figure 3, for example, PD_1 is the game between E and S_{W1} (more specifically, we will call this game PD_{E1}) and PD_2 is the game between E and S_{W2} (PD_{E2}). To simplify the notation, hereafter E's strategies in these PD_{Ej} will be C_{Ej} and D_{Ej} respectively for 'cooperation' and 'defection' where $j = S_{Wj}$.

To understand the effect of all the network's relationships on the PG played by E and S_S we start from the strategies of E and S_S in particular by investigating the process that drives the belief formation of these two agents in the PG. S_S's and E's beliefs in the PG are a function of the histories characterizing both the psychological game PG and all the PDs.

5.1. The strong stakeholder's beliefs and strategy

Player S_S's beliefs about the firm E's behaviour in the PG at time *t* depend both on the past behaviour of E in the repeated PG and on the behaviour of E in the repeated Prisoner's Dilemmas in which it is involved (in our example: PD_{E1} and PD_{E2}). The latter, because of the MG trigger strategy, is also related to all the other Prisoner's Dilemmas played in the network. Essentially, S_S forms his/her belief about E's behaviour in the PG by looking at the moves made by E in the previous periods, both in the PG and in the PD_{Ej}. In particular, before giving more technical formalization, we assume that the belief formation of S_S is based on the following considerations:

1. the initial belief of S_S is that the firm will play F_E in the PG, in consideration of the rational agreement on the CSR principle *T* subscribed to by the firm (section 4.2);

2. if at any time E does not play F in the PG, thereafter the trust of S_S in the 'fair' behaviour of E goes to zero (sufficient condition);

3. S_S's belief also depends on the moves made by E in the repeated Prisoner's Dilemmas that it plays with weak stakeholders. If the firm E always cooperates with all its weak stakeholders (i.e. it plays $C_{Ej}\ \forall j$), then the trust of S_S in the fair behaviour F_E of E remains unchanged. If at any time E defects in one repetition of a Prisoner's Dilemma that it plays with a S_{Wj}, his/her belief changes.

4. However S_S's trust in E does not change in consideration of the fact that somewhere in the network a player different from E has defected and that, owing to player's E adoption of the MG trigger strategy, E must start punishing the S_{Wj} (i.e. E's defection is aimed at punishing some other defections occurring in the network). The simple fact that E adopts its MG trigger strategy keeps it trustworthy, because it complies with a commitment intended to prevent opportunistic behaviour in the network.

The idea is that E is not trustworthy as a fair player in the PG in two cases (besides the fact that it has evidently started to play unfairly in the PG):

a) Either if E is the first player that defects against a weak stakeholder in a repetition of the Prisoner's Dilemmas it plays with them – *in fact cooperation in*

PG is aimed at producing positive output for weak stakeholders, for this reason the defection against them in the following PDs can be reasonably associated with a 'not fair' behaviour in the psychological game.

b) Or if it does not punish the defection of other agents by avoiding to implement the MG trigger strategy – *which is exactly aimed at guarantee the cooperation in the network by resort to its implicit treat of punishment.*

For this reason, S_S's belief at time *t* depends (a) on E's move at time *t-1* in the PG; (b) on E's moves in the PD_{Ej} at time *t-1* (in particular if it defects or not); (c) on the moves of all the players involved in the PDs at time *t-1* and *t-2*. In fact: (c.1) if some agent other than E defects at time *t-1*, E keeps its trustworthiness at time *t*; (c.2) if some agent defects at time *t-2*, and if at *t-1* E does not implement its part in the MG trigger strategy, this move will be considered not consistent with E's fairness, so that it will turn E into an untrustworthy player thereafter.

To give a formal description, S_S's beliefs concerning E's behaviour in the PG are settled according to the following rules (where for the purposes of this section B^t_{Ss} means 'belief at time *t* of player S_S'):

$$B^t_{Ss} = f(H^{t-1}_{PG} \times H^{t-1}_{PDE1} \times H^{t-1}_{PDE2})$$

In particular, the probabilities that E is going to play F_E or U_E in the PG according to player S_S's first-order belief are:

- $b_{Ss}(F_E) = 1$ at time *t if* at time *t-1* E plays (F_E , C_{Ej}) *and* S_S plays (e,F) in the PG
  and if

  a) at time *t-2* $\forall k, \forall i \in R_k : C_{ki}$
  or
  b) at time *t-1* $\exists k \neq E, \exists i \in R_k : D_{ki}$

- $b_{Ss}(F_E) = 0$ at time *t*, *if* at time *t-1* in the PG E plays U_E *or* S_S plays ($\neg$e);
  and if
  a) at time *t-2* $\exists k \neq E, \exists i \in R_k$ s.t. D_{ki} *and* at time *t-1* E plays (F_E, C_{Ej})
  or
  b) at time *t-1* $\exists k = E, \exists i \in R_k$ s.t. D_{ki}

Note that b_{Ss} (F_E) = 1 is compatible with the case that having learnt that at time *t-1* $\exists k \neq$ E, $\exists i \in R_k$: D_{ki}, the player E at time *t* is reacting to such information by playing (F_E, D_{Ej}). That is, S_S does not infer from condition 1.b) that player E will play (F_E, D_{Ej}) at *t*.

Given these hypotheses, the following repeated strategy by player S_S is consistent, and we assume that it is played by S_S

1. S_S starts by playing (e,F) at time *t*=1
2. $\forall t > 1$, S_S continues playing (e,F) *if*

a) at time *t-1* in PG E plays F_E *and* S_S plays (e, F)
and if

b) at time *t-1* E plays C_{Ej} in PD_{Ej} $\forall j \in R_E$ and at time *t-2* $\forall k$ and $\forall i \in R_k : C_{ki}$
or

c) at *t-1* E plays C_{Ej} in PD_{Ej} and at the *same time t-1*, $\exists k \neq$ E, $\exists i \in R_k$, s.t. D_{ki}

3. reverts to ¬e forever otherwise

where j = 1,2 are the weak stakeholders S_{Wj} linked to E; i =1,...,*m* are agents that may have relations with a generic agent in the network (normally different from E); $i \in R_k$ are the agents included in agent k's set of relations ; k =1,...,*s* are agents in the network that have a set of relations; R_k is the set of relations that characterizes agent k.

Note again that the strategy of player S_S is compatible with the hypothesis that at time *t*, when s/he continues to play (e,F), player E reverts to (F_E, D_{Ej}) if and only if at time *t-1*, $\exists k \neq$ E, $\exists i \in R_k$, s.t. D_{ki}. In other words, player S_S does not react to the information that at *t-1*, $\exists k \neq$ E, $\exists i \in R_k$, s.t. D_{ki} by immediately reverting to a sectioning strategy ¬e. In order to do so, s/he waits for at least one period, wherein player E will revert to a sanctioning strategy D_{Ej} because of the defection that occurred in some other part of the network at the time immediately before.

According to this strategy, at any *t* S_S punishes E (which means that s/he does not enter into a relation with E and plays ¬e) if (a) E defects in the PG; (b) E fails to contribute to maintaining cooperation in the network by implementing the MG trigger strategy if someone anywhere in the network defects at an immediately previous time; (c) E defects in one of the PD_{Ej} at *t-1*. However, the player S_S's reported game strategy shows more forgiveness than the standard MG trigger strategy, which if information is received about a player defecting somewhere in the network immediately requires each

player to punish its adjacent network agent as it is involved in a repeated game wherein s/he is also involved. On the contrary, in the case of player S_S, his/her modified MG trigger strategy waits for one period before the punishment starts, giving player E the chance to show whether it is consistent with its MG trigger strategy (that is to start its punishment continuation strategy with respect to the S_{Wj} as a consequence of a breach of cooperation somewhere in the network). Thus player S_S is ready to accept one-stage defection by player E, which plays (F_E,D_{Ej}), before starting the sanctioning part of his/her repeated strategy. In fact, if E defects at time *t* as a consequence of someone's else defection at time *t-1*, S_S does not anticipate its defection and continues to play (e,F) at time *t* (i.e. s/he does not punish E at time *t*), but at time *t+1* cooperation in the PG will have anyway stopped and S_S will play ¬e. This happens even though E does not have any primitive responsibility for the occurrence of defections in the Prisoner's Dilemmas. In fact, were S_S not punishing E at time *t+1* the sanction power implicit in the MG trigger strategy could not be effective. To sum up, in order to have asanction power against E, the S_S's MG trigger strategy, does not allow playing (e,F) when E effectively defects with its weak stakeholders in the Prisoner's Dilemmas, even though E's defection is the consequence of implementation of its MG trigger strategy. But it is not so harsh as to start punishing E just because someone else in the network has defected against any other agent.

5.2 The Firm's Beliefs and Strategy

Player E's beliefs are defined according to the following rules (where for the purposes of this section B_E^t means 'belief at time t of player E'):

$$B_E^t = f(H_{PG}^{t-1} \times H_{PDE1}^{t-1} \times H_{PDE2}^{t-1})$$

In particular, the probability that S_S is going to play any of his/her PG strategy according to player's E first-order beliefs is

- b_E (e,F) = 1 at time *t*, *if* at time *t-1* in the PG S_S plays (e,F) *and* E plays (F_E)
  and if
  a) at time *t-1* $\forall k, \forall i \in R_k : C_{ki}$;
  or
  b) at time *t-1* $\exists k \neq E, \exists i \in R_k$ s.t. D_{ki}

- $b_E(\neg e) = 1$ at time *t if* at time *t-1* in the PG S_S plays $\neg e$ *or* E plays (U_E)

  or

  a) at time *t-1* $\exists k = E, \exists i \in R_k$ s.t. D_{ki}

  or

  b) E plays C_{Ej} at time *t-1* and at time *t-2* $\exists k \neq E, \exists i \in R_k$ s.t. D_{ki}

- $b_E(U) = 1$ at time t *iff* S_S plays U at time *t-1*.

Note that $b_E(e,F) = 1$ does not exclude the possibility that having learnt at time *t-1* that $\exists k \neq E, \exists i \in R_k$ s.t. D_{ki} at time *t* player E is in fact playing (F_E,D_{Ej}), and hence S_S according to E may fail to predict that E is changing its strategy. Given these hypotheses on E's beliefs, the definition of the E's relevant strategy considers the role of E both in the PG and in PD_{Ej}. Hence we state that player E acts as follows:

1. E starts by playing (F_E,C_{Ej}) at time $t = 1$
2. $\forall t > 1$, E continues playing (F_E,C_{Ej}), *iff*
   a) at time *t-1* in PG S_S plays (e,F) *and* E plays F_E
      and
   b) at time *t-1* $\forall k, \forall i \in R_k : C_{ki}$
3. E reverts to (U_E,D_{Ej}) *if* at time *t-1* in PG S_S plays ($\neg e$) *or* E plays U_E
4. E reverts to (F_E,D_{Ej}) *if* at time *t-1* $\exists k \neq E, \exists i \in R_k$, s.t. D_{ki} ;
5. At $t > 2$, E reverts to (U_E,D_{Ej}) *if* at time *t-2* $\exists k \neq E, \exists i \in R_k$, s.t. D_{ki}.

We assume that E follows the MG trigger strategy with regard to all the players involved in the repeated PDs, i.e. it defects at time *t* if it knows that a defection has occurred anywhere in the network at time *t-1*. If E does not learn about defections, it continues to cooperate in the PD_{Ej}. With regard to the PG, E plays F_E as long as S_S plays (e,F), and no defection has occurred in the network, and until itself has played D_{Ej} in the PD_{Ej} at least once in order to start the sanctioning part of its strategy when someone defects in some of the PDs, but it reverts at any time $t > 1$ to (U_E,D_{Ej}) if it learns about S_S playing $\neg e$ or U – because it has no incentive to play cooperatively in the PD_{Ej} in the

absence of the psychological payoff associated with high mutual compliance with the principle *T* in PG. It also reverts to (U_E,D_{Ej}) when it is sanctioning any deviation from cooperation occurring in the network for at least two periods – since in the first period E starts sanctioning by playing (F_E,D_{Ej}).

Thus, if E starts to defect at time *t* in PD_{Ej} in order to punish agents who started to defect at time *t-1*, given the S_S strategy already described, it knows that S_S will still play (e,F) at time *t* because s/he does not want to prevent the firm's defection aimed at implementing the MG trigger strateg. Hence, at this stage, player E plays (F_E,D_{Ej}) and S_S does not anticipate this defection. Nevertheless at time *t+1*, according to his/her strategy (see section 5.1), S_S will play ¬e and E – having already defected at least once – will play (U_E,D_{Ej}) on its own. Thereafter, E will continue to play (U_E,D_{Ej}) given that S_S plays ¬e.

Note that if S_S learns at time *t-1* about a defection in the network by one or more agents other than E, s/he starts to play ¬e only at time *t+1* even if E implements its MG trigger strategy already at time *t* by using (F_E,D_{Ej}). But s/he also punishes E at time *t+1* if it does not play the MG trigger strategy at time *t* when a breach of cooperation has occurred at time *t-1*, so that at time t it has played (F_E,C_{Ej}). On the other hand, player E's modified MG trigger strategy is so conceived that it will start defecting after any information about an agent other than itself defecting in whatever part of the network by playing F in the PG but D_{Ej} in the consequent PDs. Given the delay in reaction to the same information by player S_S – or, to put it somewhat differently, given that S_S does not react immediately to such information but only to *vis a vis* defection by E in their interaction, or in the subsequent PD_{Ej} after having played F in PG – player E may profit from one period of forgiveness in which it can reap a higher payoff than would be allowed in the case of immediate sanction by S_S.

5.3 How E and the S_S play the repeated PG according to the modified MG trigger strategies

The strategies and the beliefs discussed in sections 5.1 and 5.2 define two modified versions of the MG trigger strategy by specifying how players E and S_S will act according to such a multilateral harshly sanctioning strategy with respect to the repeated play of their particular interaction, as a result of what happens in the network. This

identifies a repeated strategy profile with respect to the particular subset of players constituted by S_S and E, and hence induces the following strategy combination whereby the psychological game PG will be solved through its repeated play.

At time $t = 1$, the strategy profile in the stage-games including the move of just S_S and E is $(e,F_{Ss};F_E,C_{Ej})$ – that is, S_S enters and plays F in the PG and the firm plays F_E in the PG and cooperates in the two PDs that it plays with its S_{Wj} (because we are considering only the strategy profile characterizing how repeated games are played by S_S and E, here we disregard S_{Wj}'s choices). This state holds through all the repetitions of the PG and $DP_{E,SWj}$ stage-games until someone in the network decides to 'defect' at some time t. In this case, there are two possible deviations from the just-defined stage-games strategy profile.

1. E carries out the sanction entailed by its MG trigger strategy at time *t+1*, that is, E plays D_{Ej} in the PD_{Ej} from *t+1* onwards. S_S's in the 'fair' behaviour of E remains unchanged only for the first period *t*, and the stage-games strategy profile involving S_S and E both at time *t+1* becomes $(e,F;F_E,D_{Ej})$. However, from time *t+2* onwards, the stage-games strategy profile becomes $(\neg e;U_E,D_{Ej})$ because the MG trigger strategy of player S_S implies not preventing the defection of E for just one period if it is the consequence of E's MG trigger strategy execution, but it requires punishment of E for all the periods after it has defected once against weak stakeholders. According to its MG trigger strategy, E will continue to sanction its weak stakeholders from time *t+1* onwards, so that from time *t+2* the continuation strategy profile within this players' subset becomes $(\neg e;U_E,D_{Ej})$.

2. For some reason, Player E does not implement the MG trigger strategy at time *t+1*. In this case, S_S at time *t+2* punishes player E for not behaving so as to render effective the sanction required for implementation of the MG trigger strategies in the network. Since player E reverts to (U_E,D_{Ej}) when it learns that S_S plays $\neg e$, the resulting strategy profile from time *t+2* onward relative to the stage-game played by S_S and E is $(\neg e;U_E,D_{Ej})$ as well.

Note that E cannot avoid the decision of S_S to play $\neg e$ when someone else starts to defect in the network. Let us suppose that after someone has defected in a PD at time *t*, the firm E decides to implement its MG trigger strategy at time *t+1* in order to avoid

player S_S's sanction at *t+2*, but it also tries to avoid the S_S sanction at *t+3* by coming back cooperating by playing C_{Ej} in the PDs that it plays at time *t+2*. Under the hypothesis that the strong stakeholder S_S is adopting his/her version of the MG trigger strategy just defined in section 5.1, this attempt will be unsuccessful. In fact, once S_S learns that the firm does not contribute to punishing other agents who are continuing defection from *t+1* onwards, s/he will in any case punish E.

6. 'Sub-Game Perfection' and Endogenous Sustainability of Cooperation

The aim of the previous section was to define a modified version of MG trigger strategies by players E and S_S able to support cooperation in all games played in the network because player E is sanctioned by S_S if it defects in any PD_{Ej}. According to the analytical framework set out in section 2.1, these strategies played simultaneously with standard MG trigger strategies adopted in each repeated game by each pair of adjacent agents in the network define a repeated game Nash equilibrium (verification of the conditions for existence of this equilibrium is delayed until section 7). Our aim here is to verify whether execution of player S_S's modified MG trigger strategy is really compatible with player S_S's incentives in the iterated PG: that is, sanctioning player E is player S_S's best response if E defects in some PD_{Ej}. The relevant game theoretical concept here is the S_S strategy's 'sub-game perfection'. In other words, if the repeated play of games according to the players' MG trigger strategies (modified or otherwise) were to reach branches or sub-games out of the equilibrium path, then in that contingency the sanctions implicit in player S_S's strategy could be rationally carried out in accordance with player S_S's incentives. In this regard, we first present an intuitive analysis of sub-game perfection with reference to the stage-game PG psychological equilibrium taken as a game on its own. We will make informal use of the idea of 'trembling hand': that is, the possibility that, owing to random mistakes occurring when a given equilibrium strategy profile is played, any whatever part of the relevant game tree out of the equilibrium path can be reached, even though with low probability. By considering the possible deviations due to random mistakes, we will verify that, in any situation, player S_S's MG trigger strategy requires only playing a stage-game psychological equilibrium. In other words, even in sub-games out of the equilibrium path the strategy profile is always compatible with the principle of a player's best response.

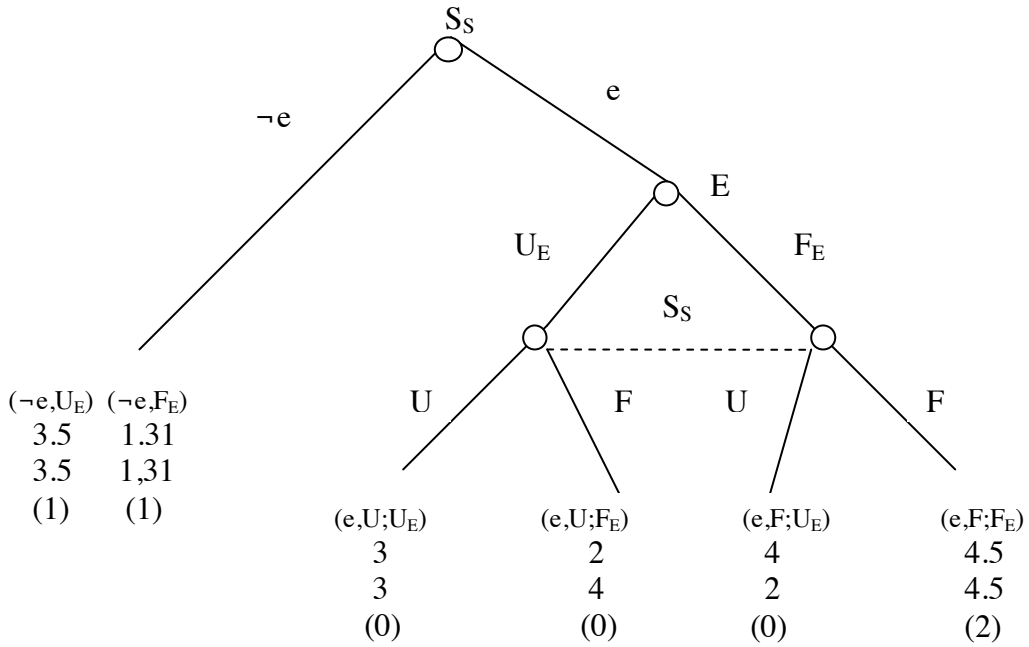
6.1 Sub-game perfection in the psychological stage-game

Before considering the sub-game perfection of the entire MG trigger strategy of players S_S and E, let us determine whether some instability (equilibrium imperfection) may be found in the psychological equilibria of the PG stage-game – considered on its own – based on player S_S's and E's conformist preferences. Figure 8 illustrates the PG in extensive form under the hypotheses of mutually consistency and common knowledge of (at least) first- and second-degree beliefs – which are typical of psychological games. Payoff vectors reported on the edge of each game tree branch show that both players have iteratively predicted that they would play the moves belonging to the path reaching a particular edge. Hence, they include the ideal component of players' payoffs that materialize when mutual beliefs are reciprocally consistent and conformist preferences are activated. To satisfy the conditions on parameters given in section 3.4 ($\lambda > d - b$ and $\lambda > c - a$) we here assume $\lambda = 2.5$. An explanation is required for the two payoff vectors reported at the branch edge $\neg e$. Each vector assigns the players' overall payoffs (included ideal utilities) based respectively on a different beliefs system concerning how the game would have been played in the remaining part of the game tree. On the left side is the psychological payoffs vector under the hypothesis that reciprocally consistent first- and second-order beliefs predict that players would choose $(\neg e, U_E)$, while on the right side is the psychological payoffs vector under the hypothesis that reciprocally consistent first- and second-order beliefs predict that players would choose $(\neg e, F_E)$.

We use intuitively the notion of sub-game perfection to analyze this game. Hence, for each psychological equilibrium, we will consider what would happen if, under the hypothesis that players are playing a particular equilibrium strategy profile, some sub-game or branch is reached out of the equilibrium-path, and whether in this case playing according to the equilibrium strategies would be irrational for the relevant player. In order to conduct this analysis we use an intuitive application of the "trembling hand" argument. Reinard Selten (1967, 1975) suggested this idea in order to introduce a random perturbation into games by means of uncorrelated small probabilities of deviation, so that, with some probability, each sub-game or branch of the game tree – also out of the equilibrium path – can be reached when players are in fact playing a given equilibrium. Equilibrium perfection consists in robustness of the equilibrium

behavior under the game perturbation induced by such small probabilities of uncorrelated random deviation "by mistake". Note that the extensive-form game of Figure 8 includes two sub-games: one starting from the second information set attributed to player E, and one starting from the individual choice attributed to player S_S at the first information set, beyond the entire game itself.

Figure 8 Extensive form of the PG stage-game with consistent belief systems and conformist preferences



To begin with, consider that players S_S and E are playing the psychological equilibrium (e,F;F_E) and are hence endowed with the relevant mutually consistent beliefs that predict such a state and consequently induce their conformist preferences. Then introduce with small probability a random mistake that when player S_S is playing e, s/he is in fact playing ¬e, so that s/he ends the game. Assuming that player E knows this random mistake probability, at the second decision node should it play differently with regard its equilibrium strategy? Consider that the players' beliefs are consistent with (e,F;F_E). Then the selection of ¬e under the belief that player E chooses F_E will entail, with small probability, a psychological payoff 1.31 for player E, which would be enhanced if, in the case of mistake, player S_S entertained the belief that player E is

playing U_E. On the other hand, if player E changes its choice to U_E at the second information set, it can fool player S_S, who – believing that E has played F_E –continues to play F, so that E reaches a payoff 4 less than 4.5. Nevertheless, consider that player S_S must know that player E has changed its behavior at the second information set because of the probability of a mistake; otherwise she would not believe that player E has chosen U_E when s/he mistakenly plays ¬e (enhancing its payoff to 4 instead of 1.31). Because of this prediction, however, s/he would play U at the third information set. As a result, in order to obtain a tiny improvement in its payoff in the case of a mistaken ¬e choice, which occurs with very small probability, with high probability E forgoes a payoff 4.5 to obtain a payoff 3 instead, which is clearly irrational. Hence, under the 'trembling hand' hypothesis, player E must not change its behavior with respect to what is required by its strategy in the equilibrium $(e,F;F_E)$.

Now consider the hypothesis that players are playing the psychological equilibrium $(\neg e;U_E)$ with reciprocally consistent beliefs. Then introduce the small probability of mistake that, when playing ¬e, player S_S is in fact playing e. They are thus allowed to reach the sub-game that starts from the second information set, which is out of the equilibrium path. Should this perturbation of the game tree induce player E to change its strategy, which requires it to implement the move U_E? Certainly not. Consider that, since they are playing according to the equilibrium $(\neg e;U_E)$, player S_S believes that player E plays U_E if its information set is reached by a random mistake occurring at the first node. Consistently with this belief, player S_S's best response is to choose U if the third information set is reached (by a random mistake). Moreover, in order to be predicted as playing U_E (according to the rationality assumption), player E must believe that player S_S, if his/her second decision node has been reached, would play U. Thus player E's best response is to play U_E at its information set if, by mistake with small probability, it is reached. This incentive-compatible behavior in the sub-game gives player E a payoff 3 that adds with small (mistake) probability to the high probability payoff 3.5. What, on the contrary, is the case if player E decides to change its move at the second decision node? Because player S_S's beliefs are still consistent with the equilibrium $(\neg e;U_E)$, she will nevertheless play U, so that E obtains a poor payoff 2 instead of 3, which is clearly irrational.

Finally, consider the case that players are playing the equilibrium (e,U;U_E) and have beliefs consistent with this psychological equilibrium – that is, they are both predicted to play U after S_S has entered. But again introduce the small random mistake probability that when S_S is playing e, she is in fact playing ¬*e*. What is player E's reaction to this probability of mistake? Consider that, under the current beliefs of the players – that is, in the sub-game they will both play (U;U_E) in the case of mistake (with small probability) – they will get a payoff 4 higher than 3. In fact, if player S_S chooses ¬e when believing that player E will play U_E (which is nevertheless required by the psychological equilibrium under consideration), then the psychological payoff is 4. Could player E enhance its payoff further by changing its behavior to F_E? Certainly not in the case of a mistake, for if S_S chooses ¬e while believing that player E is playing F_E the psychological payoff for both decreases to 1.31. Why, therefore, should player S_S believe that player E in the case of mistake is changing its behavior so that its own payoff is reduced? But if player S_S has no reason to believe that player E is changing its behavior, she will play U when his/her decision node is reached (s/he, in fact, continues to believe that E is playing U_E); hence, by changing its move, player E would worsen its payoff from 3 to 2 with high probability. Thus there is no basis for saying that incentive compatibility and the logic of best response under the perturbation hypothesis would induce the players to change their moves in the game.

To sum up, under the intuitive 'trembling hand' hypothesis that allows players to the reach any branch of the game tree out of the equilibrium path, nothing authorizes them, as long as they are rational, to make any significant modifications with respect to what is required by each of the three psychological equilibria.

6.2 Definition of the relevant sub-game

Each adjacent pair of agents in the relational network are players involved in two subsequent repeated games, except for player E, that plays three repeated games with its adjacent stakeholders, and S_S who plays just one repeated game with E. The strategies whereby all players make their choices in each stage-game at any time are made conditional on choices made by all other players in the network through the assumption that each player adopts an MG trigger strategy (including the modified version defined in section 5). These are rules for deciding how to play any stage-game at any time in function of the past history of the game. However, MG trigger strategies have the

peculiarity that how each player chooses at any time *t* in a given stage-game depends on the decisions made at a time *t-1* by any other player participating in the network, also playing a different and remote repeated game. In fact, if a defection occurs somewhere in the network, any player, according to his/her MG trigger strategy, starts to punish the players s/he is related with, thereby changing any player's incentive to continue cooperation in the immediately subsequent game that s/he plays with his/her successor in the network. This construction makes it possible to consider all the stage-games played at time *t* as if they were sub-games of a unique dynamic game played at any time *t* by all the network's agents. Moreover, the dynamic game is repeated ad infinitum, and the way in which each repetition is played – under our current assumptions – is dictated at any time by the players' MG trigger strategies.

Within this context, we must define the proper sub-game to be analyzed. It is necessary to select a sub-game that may convey not just the information that E has abandoned its stage-game equilibrium strategy F_E, shifting to the other stage-game strategy U_E, but also the information that, in some subsequent PD games with S_{Wj}, it has played D_{Ej} instead of C_{Ej} after having played the strategy F_E in PG . Put differently, it is necessary that the stage game – taken as the relevant sub-game of the overall dynamic game played by all the network's players – allows player S_S to entertain correct beliefs not only on the choices F_E or U_E that player E makes in the PG, but also on choices that it makes in the subsequent PD_{Ej}. Of course, player S_S needs to understand whether player E is consistent with a 'fair' mode of playing the PG (strategy F_E of the stage game) also in consideration of how it plays the following PD_{Ej} game, because it is only in these games that the amount of surplus saved on behalf of players S_{Wj} will be effectively allocated to pay them fairly for their cooperation with E. Recall that this was our first assumption in section 4.3 and that it was also incorporated in the assumption that player S_S believes that player E is playing F_E with probability zero if s/he learns about its defection in the subsequent PD_{Ej}.

The underlying intuitive idea is that if in the PG stage-game one or both of the players choose the collusive and egoist strategy U, no part of the surplus is saved or entitled to S_{Wj}, so that the result of PG has no effect on the payoffs accruing to the S_{Wj} in the subsequent PD_{Ej} games. This is clear when S_S plays U unilaterally, since s/he simply takes away from the game for his/her personal consumption the extra-rent that

could be allocated to the S_{Wj} for his/her personal consumption. But this is also true if E chooses U_E because, for instance, E thus allocates to the private earnings of E's shareholders or managers any extra-rent that otherwise could be an endowment available to the firm in order to improve its cooperation with S_{Wj}. Thus, if the players choose U or U_E in PG there is no information that can arise from the subsequent games concerning player E's consistency with the adopted strategy or the effective payoffs engendered in PG. In these cases player S_S will obtain directly from the equilibrium solution of PG all the information necessary to establish that E plays unfairly, so that s/he will anyway not trust E for 'Fair play'. Choices like C_{Ej} and D_{Ej} in these cases may only give information about how player E responds to 'external' incentives (with respect to PG) deriving from the subsequent stage-games or the MG trigger strategies that players adopt to play these repeated games and are indifferent with respect to the PG game payoffs. If these choices are reported in the sub-game under consideration it is only for completeness of the formal representation, and without giving any information about their outcome in the subsequent games. Their attached payoffs are only relative to the PG, with respect to which they are indifferent. To be sure, nor does the information concerning the choice of PD_{Ej} strategies by player E if the PG was played according the equilibrium $(e,F;F_E)$ give any information about the payoffs distribution depending on the solution of subsequent PD_{Ej} games. What it does provide, however, is very relevant information concerning whether the PG payoffs really correspond to what is expected from playing the equilibrium $(e,F;F_E)$.

In fact, when the PG is played according to the equilibrium $(e,F;F_E)$ a part of the surplus is saved and entitled to the S_{Wj} (according to Figure 5 it amounts to 2 utils). The interpretation is that player E is committed to using it in order to pay the S_{Wj} a fairer payoff for mutual cooperation in the PD_{Ej} games. This will not change – as we shall soon see – the basic strategic structure of the PD_{Ej} game. It can be considered as only an addition to the payoff that S_{Wj} gets conditionally on how player E will play these games. In particular, if player E chooses to cooperate by C_{Ej} with the S_{Wj}, the amount of 2 utils saved on behalf of S_{Wj} is effectively used to pay him/her more than the standard PD_{Ej} payoffs otherwise characterizing E's relations with weak stakeholders.

Figure 10(a) The basic PD_{Ej} in normal form

E \ S_{Wj}	C_{jE}	D_{jE}
C_{Ej}	2, 1	-1, 2
D_{Ej}	4, -1	1, 0

Figure 10(b) The PD_{Ej} if the antecedent PG has been solved by $(e,F;F_E)$

E(e,F, F) \ S_{Wj}	C_{jE}	D_{jE}
C_{Ej}	2.5, 2.5	-1, 4
D_{Ej}	6.5, -1.5	1, 0

To illustrate how the PG game equilibrium solution $(e,F;F_E)$ may affect the subsequent PD_{Wj}'s payoff levels, see Figures 10(a) and Figure 10(b). The first figure is a numerical example of the basic PD game played by any two adjacent players in the network. It also represents the interaction between E and S_{Wj} seen as independent from the conclusion of the antecedent game played by E and S_S. The figure reports the PD_{Ej} game as it will typically unfold if the antecedent PG game had an unfair solution such as $(e,U;U_E)$, or $(\neg e;U_E)$. The second figure illustrates how the former payoff matrix is changed by the additional payoffs 2 provided to S_{Wj} by the solution $(e,F;F_E)$ reached by E and S_S in the antecedent PG, granted that E plays cooperation C_{Ej} in the PD_{Ej}. Note, however, that in PD_{Ej} player E is not constrained to do so by the solution of the antecedent game PG, since it can choose its strategy freely, and also appropriate the extra-rent by playing D_{Ej} in the game.

The payoff transformation in 10(b) can be explained as follows. The endowment of 2 utils saved on behalf of player S_{Wj} through the fair solution of the antecedent PG game, is managed by player E in PD_{Ej} so that it can be mutually advantageous in the case of full cooperation between them. E allocates the endowment to paying player S_{Wj} a higher wage in exchange for a player's S_{Wj} extra-effort with respect to what was already incorporated in payoffs of Figure 10(a). Effort enters S_{Wj} payoffs negatively (-0.5) but

produces an advantage (+0.5) for E. The result is an effectively fairer (equal) payoff in the case of mutual cooperation (C_{Ej}, C_{jE}) = (2.5, 2.5) in the DP_{Ej} (with a significant improvement of S_{Wj} payoffs with respect to the basic game). However, the game has not changed its basic Prisoner's Dilemma structure. By playing 'defection', S_{Wj} can take the entire payment (the basic wage 2 plus the additional payoff 2) without incurring any production cost. On the other hand, if S_{Wj} agrees to increase his/her investment by 0.5, player E may appropriate the entire surplus engendered both player S_{Wj}'s basic and additional investments (4+0.5) plus the additional 2 utils that were saved on behalf of S_{Wj}, but in this case were in fact simply 'robbed' by E.

It can also be verified that the payoff transformation by means of the additional 2 utils does not change the players' incentive to cooperate in the *repeated* PD. In particular, it does not eliminate the basic asymmetry that characterizes the PD_{Ej}. That is to say, whereas each S_{Wj} considers continuous cooperation with E worth carrying out by repeated plays of the game, player E (the firm) does not find it sufficiently profitable to play iterated cooperation with the S_{Wi}, and prefers to defect even in the repeated game. This can be seen by comparing the critical discount rates δ^* that make repeated cooperation for the two players profitable under the two cases with their actual discount rate δ. In the basic and modified case respectively, the player's E critical discount rates are

$$\delta_E^* = (4-2)/(4-1) = 0.666, \qquad \delta_E^{**} = (6.5-2.5) / (6.5 -1) = 0.7272$$

Since by assumption player E's actual discount rate (or level of myopia) is $\delta < \delta^*$, it is necessarily also $\delta < \delta^{**}$ (since 0.666 < 0.7272), so that in the modified PD_{Ej} game E has an even more intense incentive to defect from repeated cooperation. On the other hand, the respective critical discount rates that make repeated cooperation profitable for players S_{Wj} in the two cases are

$$\delta_{Swj}^* = (2-1)/2 = 0.5, \qquad \delta_{Swj}^{**} = (4-2.5)/4 = 0.375$$

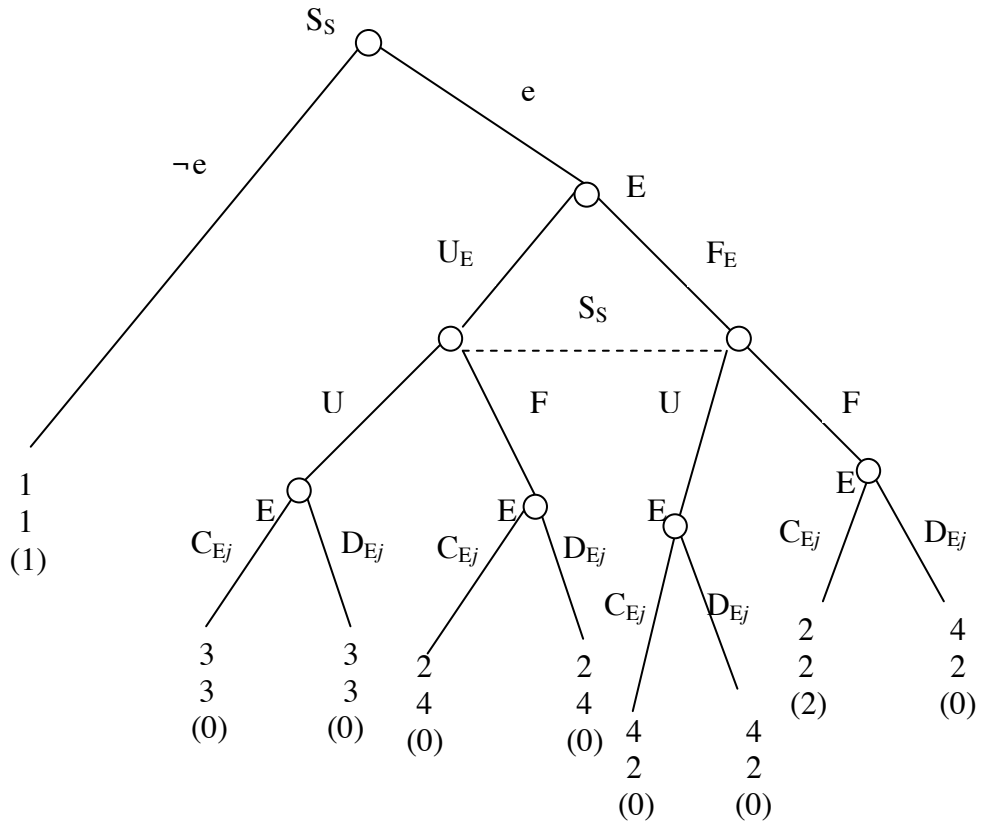
In this case, by assumption player S_{Wj}'s actual discount rate (or myopia level) is $\delta > \delta^*$ and hence necessarily $\delta > \delta^{**}$ (since 0.5 >0.375). Whereas the payoff-transformed PD_{Ej} game – due to the antecedent PG game's fair solution – makes players S_{Wj} even more willing to engage in mutually profitable cooperation with E, nonetheless the transformed PD_{Ej} reinforces game player E's preference for defection. Therefore the

external support for cooperation deriving from the 'Fair play' psychological payoff in the PG is even more important in order to sustain cooperation in the PD_{Ej}

There is consequently a very compelling sense in which player S_S needs to assess player E's behavior in the subsequent PD_{Ej} in order to ascertain whether the fair strategy F_E has been effectively played in PG. To understand whether player E has effectively implemented the strategy F_E chosen in PG, s/he must check E's behavior until the subsequent stage game is reached, wherein the allocation of the endowment to improve S_{Wj} conditions is carried out through C_{Ej}. Otherwise, the F_E choice in PG would be ineffectual or simply apparent, since what in fact results is the same outcome that E could have determined by choosing U_E when S_S chose F (i.e. E appropriates the residual of 2 utils set aside by S_S). In this case, player S_S considers the player E's pair of subsequent moves (F_E,D_{Ej}) as essentially identical to playing U_E in the PG (recall S_S's learning rule in section 5.1).

Consequently, the relevant sub-game must include the following information: has the strategy adopted by E in the subsequent PD_{Ej} effectively allocated the payoff 2 to the dummy player according to the saving decisions (F;F_E)? If E plays C_{Ej} it has effectively implemented the strategy F understood as consistent with the *T* principle agreed in the pre-play stage of PG. If E plays D_{Ej} it has simply betrayed player S_S. The proper sub-game is given in Figure 11.

Figure 11 The relevant sub-game in extensive form, illustrated only in terms of material payoffs



Note again that, in order to convey the relevant information, the sub-game includes the choices C_{Ej} and D_{Ej} in the subsequent PD_{Ej} but does not anticipate the description of the following stage-game payoffs. However, if player E adopts the strategy (F_E, D_{Ej}), against S_S playing (e,F), the material payoff vector becomes (2,4,0). The psychological payoffs change accordingly. Only if player E plays the pair (F_E,C_{Ej}) when S_S plays (e,F) are the material payoffs of the PG (2,2,2), which may give rise to a psychological equilibrium of the game.

As in section 6.1, the psychological payoffs can be computed under the assumption of mutually consistent and common knowledge of reciprocal first- and second-order beliefs that activate conformist preferences (once again it is assumed that $\lambda = 2.5$). Figure 12 illustrates the corresponding sub-game in normal form, where the psychological payoffs are computed to represent conformist preferences.

Figure 12 Normal form of the relevant psychological sub-game

S_S \ E	(F_E,C_{Ej})	(F_E,D_{Ej})	(U_E,C_{Ej})	(U_E,D_{Ej})
(e,F)	4.5, 4.5, (2)	2, 4, (0)	2, 4, (0)	2, 4, (0)
(e,U)	4, 2, (0)	4, 2, (0)	3, 3, (0)	3, 3, (0)
¬e	1.31, 1.31, (1)	3.5, 3.5, (1)	3.5, 3.5, (1)	3.5, 3.5, (1)

Player E's strategies are labeled C_{Ej} and D_{Ej} only in order to account for what may happen in the stage-game PG because of these components of player E's strategies as well. Again, no consideration is given here to the payoffs that these strategies will accrue to player E when the proper PD_{Ej} is played. Recall also that only when they are associated with F_E are the strategies C_{Ej} and D_{Ej} material to this sub-game. Inspection of the psychological payoff matrix shows that the three psychological equilibria present in the game of Figure 6 and discussed in sections 3.3 and 3.4 also exist in the just-defined sub-game.

Consider first the stage-game strategy profile $(e,F;F_E,C_{Ej})$. This is the sub-game psychological equilibrium inducing 'Fair play' in the PG and 'cooperation' by player E in the subsequent PD_{Ej}. In fact, the chosen value of λ and mutually consistent first- and second-order reciprocal beliefs predicting that player S_S will use (e,F) and player E will use (F_E,C_{Ej}), respectively, induce the psychological payoffs vector (4.5,4.5) for the two active players that makes such strategies clearly mutual best responses. The distinctive feature of this sub-game representation is that, in order to give rise to such a 'Fair play' psychological equilibrium, player E's consistency in the consequent PD_{Ej} game must be included in the strategy description. This consists in using the cooperative strategy C_{Ej} that entails no appropriation by E of the surplus share saved for S_{Wi} by choosing the Fair strategies F and F_E in PG.

Also the strategy profile $(\neg e;U_E,D_{Ej})$ is a sub-game psychological equilibrium. If both the players reciprocally believe that E will play U_E in the sub-game if S_S enters, whereas player S_S will 'stay out' by playing ¬e, given the chosen value of λ the payoff

vector in the sub-game for the two active player becomes (3.5,3.5), and ¬e, (U_E, D_{Ej}) are the mutual best responses. This equilibrium is apparently weak because E has two further strategies, (F_E,D_{Ej}) and (U_E,C_{Ej}), that give the same psychological payoffs when player S_S chooses ¬e and beliefs are aligned with the relevant strategy profiles. But this is not the case. To see why, consider the third strategy profile (e,U;U_E,D_{Ej}). If the beliefs of players E and S_S are such that each thinks that they will play collusively and that they believe that she/it will play collusively, then the value of λ goes to 0 and the payoff vector for active players is (3, 3), which entails that (U;U_E) is a pair of mutual best responses in the sub-game. Clearly, this is a strategy profile that defines a psychological equilibrium in the sub-game under consideration, and also in all the subsequent PD_{Ej} – where it coincides with the unique equilibrium point of one-shot Prisoners' Dilemmas. Recall in fact that player E's strategy (U_E,D_{Ej}) means that it will defect in its relationship with weak stakeholders in the PD_{Ej}, which is in line with E's incentives internal to the subsequent Prisoners' Dilemma Games seen as sub-games (so that there is no difficulty in maintaining that any strategy profile of the current sub-game that prescribes that this player E strategy choice will be incentive-compatible in the following sub-games for E).

Note the importance of the foregoing argument in regard to the apparent weakness of the sub-game equilibrium (¬e;U_E,D_{Ej}). The strategy (U_E,D_{Ej}) is compatible with both the last two equilibria, and for whatever mutually consistent belief system, in at least one case (U_E,D_{Ej}) gives player E an higher psychological payoff than (F_E,D_{Ej}). Therefore it is weakly dominant on the strategy (F_E,D_{Ej}). Since weakly-dominated strategies like (F_E,D_{Ej}) can be eliminated, there is no reason for E to be consistently believed to have chosen (F_E,D_{Ej}). Thus the strategy profile (¬e;F_E,D_{Ej}) is not a reasonable psychological equilibrium of the sub-game (the mutually consistent belief system that could justify it is not consistent with common knowledge of rationality). But what about E's strategy (U_E,C_{Ej})? Under the proper beliefs systems this allows strategy profiles (e,U;U_E,C_{Ej}) and (¬e; U_E,C_{Ej}) that correspond to payoff vectors (3, 3) and (3.5, 3.5). These are psychological equilibria of the sub-game, so that (e,U;U_E,D_{Ej}) also seems to be a weak equilibrium, while (¬e;U_E,D_{Ej}) remains weak owing to this indifferent alternative. But consider that (U_E,C_{Ej}) entails that player E will cooperate in the subsequent PD_{Ej} games, under the conditions that in the antecedent PG game the equilibrium solutions are either (e,U; U_E, C_{Ej}) or (¬e; U_E,C_{Ej}). Both such profiles

exclude 'Fair play' in the PG and do not provide E with any conditional incentive for cooperation in the PD_{Ej} (recall that player E's MG trigger strategy requires it to play 'defect' in the subsequent DP_{Wj} if the antecedent PG game player S_S's strategy has been either (e,U) or ¬e). Thus the strategy (U_E,C_{Ej}) is clearly dominated by the alternative (U_E,D_{Ej}) in the sub-games that follow the one considered here, and hence cannot be considered as part of reasonable psychological equilibria of the sub-game under consideration (there is no basis for a mutually consistent system of beliefs that predicts player E will cooperate in the PD_{Ej} when S_S does not resort to a strategy that benefits E with the psychological payoffs associated with Fair play conditional on the prosecution of cooperation). Not only can player E's strategy (F_E,D_{Ej}) be eliminated in the current sub-game, but also the strategy (U_E,C_{Ej}), because it is dominated in the subsequent PD_{Ej} sub-games – being not superior to (U_E,D_{Ej}) in the current one. Consequently, there are only three strategy profiles that are reasonable psychological equilibria in the sub-game.

6.3 Sub-game perfection of players' S_S and E MG trigger strategies

In this section we finally show that the combination of player S_S and E's modified MG trigger strategies as defined in sections 5.1, 5.2 and 5.3 is a sub-game perfect equilibrium. The task is accomplished by considering various cases in which one can observe a deviation from the equilibrium path that would be traced in the current sub-game under the hypothesis that the two players follow their MG trigger strategies. We will verify the equilibrium property of choices that the players should make according to this pair of repeated strategies out-of-the-equilibrium-path in the relevant sub-game. This again employs an intuitive version of the "trembling hand" argument used in section 6.1.

To begin with, recall that execution of S_S's and E's pair of MG trigger strategies, adopted to play repeated games, entails in the sub-game currently under consideration that the strategy profile $(e,F;F_E,C_{Ej})$ will be implemented. Thus, as long as neither player deviates from his/her equilibrium strategy, this strategy profile induces 'Fair play' in each repetition of the PG and player E's 'cooperation' in each repetition of any PD_{Ej}. The learning rules whereby the players adapt their beliefs to the past behaviour of players in the network work as stated in sections 51 and 5.2 respectively. Finally, if player S_S understands that player E is *de facto* playing U_E in the PG, his/her MG trigger

strategy dictates reverting to ¬e. At the same time, when player E learns that player S_S will not keep playing (e,F) but from the foregoing period has changed to U or ¬e, according to its MG trigger strategy, it must also change to U in the PG and also to D_{Ej} in the subsequent PD_{Ej}.

Hence, assume that when players E and S_S are adopting the modified MG trigger strategies, there is a small probability of the occurrence of a random mistake such that at time t they find themselves out-of-the-equilibrium-path. According to the sanctioning part of their grim trigger strategy, actions would produce a strategy profile different from $(e,F;F_E,C_{Ej})$ in the current sub-game (see Figures 11 and 12). The random deviation is imputable to player E because of one of three possible mistakes: (a) at time *t-1*, contrary to expectations, E has stopped playing F_E and started to play U_E in the PG; (b) at time *t-1*, after playing F_E as expected, E has been the first in the network to play D_{Ej} (without any justification); (c) at time *t-1*, after the information was transmitted throughout the network that one member had played uncooperatively at time *t-2*, E continued playing C_{Ej}.

According to his/her learning rules, after having observed at time *t-1* U_E or (F_E,D_{Ej}) or also (F_E,C_{Ej}) (in the special case that information circulated that someone else had played D in some PD in the network), at time *t*, S_S realizes that player E is *de facto* playing the PG unfairly, that is, the probability of F_E is 0. Thus his/her MG trigger strategy requires that S_S play ¬e at time *t* in the PG (which is coherent with these beliefs).

On the other hand, player E's MG trigger strategy requires it to play (U_E,D_{Ej}) because the condition for continuing to play (F_E,C_{Ej}) that nobody in any PD at time *t-1* deviated from C_{ki} has been violated either by E itself (case b) or by another player in the network (case c). In fact, for these cases, E's learning rules state that the probability of S_S playing ¬e is 1. Moreover, player E's MG trigger strategy requires it to start playing (U_E, D_{Ej}) if E itself at time *t-1* played U_E (coherently with its learning rule that predicts in this case that the probability of S_S playing (e,F) is 0).

Do these strategies induce any irrational choice in the relevant sub game out-of-the equilibrium-path? Note that if E plays (F_E,D_{Ej}) at *t-1* (case b), it would not be rational for E to continue playing in this way, because this is a weakly dominated strategy. If E thinks that S_S is going to play ¬e at *t*, (F_E,D_{Ej}), it would not be a better response than

(U_E,D_{Ej}). But if E realises that S_S thinks that it believes s/he is going to play U, so that she chooses U, then playing (F_E,D_{Ej}) at *t* would be inferior to playing (U_E,D_{Ej}). There is no reason for E to play a strategy that can only give it less than the alternative. This is consistent with player S_S's learning rule that induces him/her, after observing at *t- 1*, to believe that E will play U_E. Thus the profile (e,F;F_E,D_{Ej}) can be only a transitory state from the initial profile to a different continuation strategy profile. It cannot stabilize. Neither could player E respond to the deviation by playing in the sub-game (even though this was its deviation at *t-1*). In fact, player S_S's learning rule induces him/her to play ¬e and it must be believed by E. Moreover, there is no incentive for E's repeated cooperation in the subsequent PD_{Ej} without a psychological payoff deriving from PG. Finally, in the cases of both mistake b) and c), E must know that throughout the network players have started the sanction stage of their MG trigger strategies, so that there will no longer be cooperation in the PD(s). Thus replying the deviation by (U_E,C_{Ej}) would be irrational. By contrast, the profile (¬e;U_E,D_{Ej}) is a psychological equilibrium of the sub-game, and under the appropriate mutually consistent reciprocal beliefs system it could emerge as a completely rational combination of mutually best responses. Indeed, player S_S's rules of belief adaptation predict that E will play U_E, while player E's rules of belief adaptation predict that player S_S will play (¬e). These beliefs are common knowledge. Thus each player has a second-order belief predicting exactly the change of beliefs which is occurring to the other player. Given the first- and second-order beliefs that they will play the pair (¬e; U_E) in the PG at t, player S_S must also believe that E will play D_{Ej} in the subsequent games, and this is also player E's only second-order belief about S_S's beliefs that is consistent with E's choice. Under our assumption of the value of λ, conformist preferences are activated in the PG and the psychological equilibrium (¬e; U_E, D_{Ej}) arises at time t in the sub-game. The deviation from the equilibrium path – after one stage – induces the transition from one psychological equilibrium of the sub-game to another. The strategy profile in which S_S sanctions the deviation coincides with the emergence of a sub-game psychological equilibrium, so that there is no instability in the required behavior and the carrying out of the threat is perfectly credible.

But now assume that the relevant deviation in E's behavior occurs at time *t-1* because of a choice by a S_{Wj} player who – in contrast with the execution of his/her MG trigger strategy – during the cooperative stage *t-2* mistakenly deviates to C_{Wj}. According

to its MG trigger strategy at *t-1*, player E must play (F_E,D_{Ej}), and at the subsequent time *t* it must play (U_E,D_{Ej}). The deviation at time *t-2* does not immediately affect S_S's behaviour in the sub-game at time *t-1*, because his/her beliefs about E change only conditionally on the learning of its effective choice in a stage-sub-game. Thus, in the transition stage *t-1*, S_S still chooses (e,F), while player E chooses (F_E,D_{Ej}) , giving rise to (e,F; F_E,D_{Ej}). This is clearly an unstable strategy profile that may last only the time necessary for player S_S to realize that E is *de facto* playing the sub-game unfairly. From time *t* onwards, the players will revert to the sub-game psychological equilibrium (¬e;U_E,D_{Ej}) through a line of reasoning completely analogous to the one given for deviations directly due to player E's mistakes. Essentially, at time *t-1*, E correctly does not changes its beliefs about S_S since it knows that his/her learning rules and strategy forgives a single period in which E may play (F_E,D_{Ej}) in order to start punishing S_{Wj}. From t onwards, however, player S_S's first-order beliefs will be aligned with player E's behaviour, and also player E's beliefs about S_S's choice ¬e and their mutual second-order beliefs are aligned. The sub-game psychological equilibrium (¬e; U_E, D_{Ej}) again emerges – which is consistent with the sanctioning stages dictated by the players' MG trigger strategies.

Finally, a deviation may also arise from a mistake by player S_S. At time *t-2*, player S_S chooses U, in contrast with his/her MG trigger strategy, while player E still chooses (F_E,C_{Ej}). The result in the sub-game at time *t-2* is a disequilibrium transition state (U;F_E,C_{Ej}). Players do not have mutually consistent beliefs, since – to exemplify – E fails to predict S_S's choice, believing mistakenly that s/he is still choosing (e,F), and S_S believes that E fails to predict his/her behavior because E believes it is still (e,F) when s/he is choosing U instead.

At time *t-1*, because of the rule of beliefs adaptation, player E comes to believe that S_S chooses U with probability 1, and in the relevant sub-game, owing to its MG trigger strategy, E starts playing (U_E,D_{Ej}). At the same time, S_S correctly believes that E is playing (U_E,D_{Ej}), because of the learning rule whereby s/he no longer believes at *t* that E will play F_E if some player deviated at time *t-2* from its component of the strategy profile (e,F;F_E,C_{Ej}). Moreover, because S_S knows that it is unprofitable for player E to cooperate in the iterated PD_{Ej} when there is no Fair play in the PG, S_S also predicts D_{Ej}. Because of common knowledge of the players' beliefs adaptation rules, it is likely that,

at *t-1*, players entertain the following second-order beliefs: player E predicts that S_S believes it is playing (U_E,D_{Ej}); player S_S predicts that E believes s/he is choosing U.

Thus, if S_S were effectively choosing U at time *t-1*, the result would be (e,U;U_E,D_{Ej}). Given the aforesaid first- and second-order beliefs – b_E =U, b_{Ss} = (U_E,D_{Ej}) ; b_E^2 = (U_E,D_{Ej}), b_{Ss}^2 = U – that strategy combination would be a psychological equilibrium of the sub-game: to be sure, a psychological equilibrium wherein the players' ideal payoffs are nil, because of the unfair distribution, but nevertheless a psychological equilibrium that would stabilize and replicate at time t and thereafter. This would entail that, when a random deviation is caused by S_S, a collusion equilibrium is reached in the sub-game at time *t-1*, contrary to the requirements of the MG trigger strategies of both players, which command that any deviation be sanctioned by the stay-out strategy of player S_S.

However, this is not the case. It is true that player S_S's adaptation rule states that if s/he at *t-2* has not chosen (e,F), then s/he believes with probability 1 that at time *t-1* E will do U, but his/her modified MG trigger strategy also states that if at *t-2* any whatever player has deviated from his/her component of the strategy profile (e,F;F_E,C_{Ej}) then at *t-1* S_S will move to ¬e. Thus, at *t-1*, the result is in fact (¬e;U_E,D_{Ej}), which contradicts player E's first-order belief that S_S does (e,U) and entails that player S_S's second-order belief that E believes that s/he does (e,U) mistakenly predicts his/her own behavior so that s/he knows that the beliefs system is inconsistent. At time *t-1* the players' reciprocal beliefs system does not exhibit the typical mutual consistency and alignment with actual behavior required for psychological equilibria. Therefore, at *t-1*, neither the psychological equilibrium (e,U,U_E,D_{Ej}) – which is what player E mistakenly predicts will happen – nor the psychological equilibrium (¬e;U_E,D_{Ej}) – which is what actually occurs, even though it is not consistently represented through the players' beliefs – emerge. In the actual state of affairs (¬e;U_E,D_{Ej}), in fact, the players cannot profit from any psychological payoff, given their mutually inconsistent beliefs system (E does not believe what S_S really does, and S_S predicts that E does not believe what s/he really does), so that they obtain only the material payoffs (1,1).

But, at time *t*, E's beliefs are finally aligned with S_S actual behavior. Because of what has been observed at *t-1*, E believes that S_S does ¬e, while S_S continues to believe that E chooses (U_E,D_{Ej}). Since they know the reciprocal rules of adaptation, they also correctly believe what they believe, and all these beliefs converging on the state (¬e;U_E,D_{Ej}) are aligned with their actual choices. This is therefore a psychological

equilibrium of the sub-game, which may stabilize and can be replicated thereafter. Moreover, it is completely consistent with the dictates of the repeated-games MG trigger strategies of the two players.

To sum up, if a player S_S random mistake occurs, two transition periods are needed before a psychological equilibrium of the sub-game is reached. At *t-2* the outcome is $(e,U;U_E,D_{Ej})$, with a worse material payoff for E, and a material advantage for S_S, no payoff to the dummy S_{Wj} and no ideal utilities at all. At *t-1* the outcome is $(\neg e;U_E, D_{Ej})$ which is not even a psychological equilibrium because of the still inconsistent players' beliefs, so that they merely obtain the 'stay-out' material payoff (1,1,1). But at *t* the psychological equilibrium $(\neg e;U_E,D_{Ej})$ is finally reached because it is supported by the appropriate reciprocal and consistent beliefs and provides psychological motivations for implementation of player S_S's sanction and support for the 'would-be-ready-to-cooperate' preference by E.

The conclusion is that, for whatever random mistake that takes the sub-game play out of the equilibrium path established by the pair of modified repeated MG trigger strategies of player S_S and E, there is no reason to think that the out-of-the-equilibrium-path choices will stabilize on a sub-game psychological equilibrium that would induce stable deviation from what the pair of modified grim trigger strategies would require the players to do. Especially, there is no reason to think that the logic and incentives faced in the sub-game will prevent player S_S from carrying out the punishment stages of his/her repeated MG trigger strategy which is at the basis of the sustainability of fair cooperation throughout the network when player E has no direct material incentive to play cooperatively with both its S_{Wj}. By contrast, after a maximum of two transition stages, a sub-game psychological equilibrium is reached which guarantees that the punishment stages of player S_S's MG trigger strategy will be implemented in accordance with his/her psychological 'incentives' and the sub-game best response logic. Assuming that the pair of modified MG trigger strategies, together with the standard ones played by any other player in the network, constitutes a repeated games Nash equilibrium, this result ensures that cooperation in the firm-stakeholders-other-agents bilaterally deficient relational network is endogenously stable (Q*uod Erat Demostrandum*).

7. Conditions for a firm-stakeholders network fair-cooperative equilibrium

This section is concerned with the precise conditions whereby the repeated games strategies of players S_S and E studied so far are a Nash equilibrium of the games that they repeatedly play between themselves and (in the case of E) in relation to weak stakeholders S_{Wj}. This has been presumed thus far in accordance with intuition and standard results concerning the MG trigger strategies used in this kind of bilaterally deficient relational network, wherein adjacent players are involved in repeated PD(s) (see section 2). We have concentrated largely on the effective sustainability of the cooperation induced by these equilibrium repeated game strategies, because the main challenge was their sub-game perfection in the stage-game wherein S_S must back all the sanctioning mechanisms without apparently having any incentive to do so in the event that the need to implement the threat of her strategy arises. But we must now show that the modified MG trigger strategies that players S_S and E use in their repeated games (the PG and PD_{Ej}) satisfy the conditions for the existence of a repeated game Nash equilibrium.

We must verify the following (the payoffs reported for the reader's convenience in Figure 13 are the same as in the PG of Figure 7):

1. S_S prefers to continue:
   - ✓ to play (e,F) instead of playing (¬e) as long as E plays (F_E) in the PG and
   - ✓ to play (e,F) instead of playing (e,U) as long as E plays (F_E) in the PG.
2. E does not have an incentive to defect either in the PG or in the PD_{Ej} as long as:
   - ✓ all the players involved in the PDs are cooperating and
   - ✓ S_S is playing (e,F)

Figure 13 Again the PG in normal form

E S_S	F_E	U_E
e,F	$b+\lambda_{Ss}$, $b+\lambda_E$, *(b)*	*b, d, (0)*
e,U	*d, b, (0)*	*c, c, (0)*
¬e	$a+1/x\ \lambda_{Ss}$, $a+1/x\ \lambda_E$, *(a)*	$a+\lambda_{Ss}$, $a+\lambda_E$, *(a)*

where $d > c > b > a$ and where the conditions for the existence of the psychological equilibria are: $b+\lambda_E > d$, $a+\lambda_E < b+\lambda_E$, $a+\lambda_{Ss} > c$.

Let us start with point 1 and consider the payoff that S_S may obtain in the repeated PG. In order to verify whether S_S has any incentive to defect and stop playing (e, F) as long as E plays (F_E), we have to compare the repeated payoff obtained by S_S when s/he plays (e,F) and E plays (F_E) with:

a) the payoff obtained by S_S when s/he plays (¬e) and consequently in the continuation of the game E plays (U_E)

b) the payoff obtained by S_S when s/he plays (e,U) and consequently the continuation of the game E plays (U_E).

If S_S and E play F and F_E respectively, S_S obtains a payoff (hereafter also the 'cooperative payoff') equal to $\sum_{n=0}^{\infty}(b+\lambda_{Stks})\delta^n/(1-\delta)$

The payoff obtained by S_S in case (a), is obviously lower than the 'cooperative payoff' because it is equal to $[a+(1/x)\lambda]\delta$ (the payoff obtained at the first stage when S_S defects (¬e) and E plays F_E) plus $\sum_{n=0}^{\infty}(a+\lambda_{Stks})\delta^n/(1-\delta)$, which is the payoff obtained by S_S from the second stage, after his/her defection, onwards (recall that $b > a$).

The payoff obtained by S_S in case (b) is:

i. *d* in the 'first' period of deviation, when S_S defects and plays (e,U) while E plays F_E;

ii. *c* from the 'second' period after the deviation onwards when the continuation profile becomes (e, U;U_E).

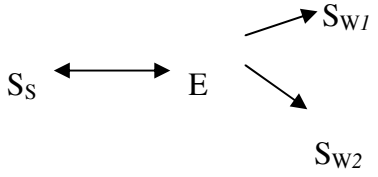
Obviously, neither is this strategy not convenient for S_S, at least if we assume that the players are endowed with high environmental cognitive social capital so that $\lambda > (d - b)$, the 'cooperative payoff', is higher than the payoff obtained by playing (e,U):

$$\sum_{i=1}^{\infty}(b+\lambda) \succ d+\sum_{i=2}^{\infty}c\,.$$

With respect to point 2 – the firm's incentive to depart from the 'fair-cooperative equilibrium' amounts to choosing 'Fair play' in the repeated PG and 'cooperation' in

the repeated PD_{Ej} – we shall consider the sub-network of relations involving E (see Figure 14).

Figure 14 The restricted firm-stakeholders network



With regard to the relation in which E is involved, note that $\sum_{j \in RE} g_{Ej} \geq 0$ is a necessary condition in order for E to continue to play F in the PG. It amounts to saying that $g_{ESs} - (g_{ESw1} + g_{ESw2}) \geq 0$.[15]

We want to show that E has no incentive to defect when it, S_S, and all the other players in the network are cooperating. E may defect by adopting two strategies.

A) E stops cooperating with S_{Wj} at time *t* and, at the same time, it continues to play F_E in the PG. Given player S_S's belief formation rule, since E is the first to defect in PD_{Ej}, S_S believes that E will defect also in PG at time *t+1*. For this reason (following his/her MG trigger strategy), S_S will punish E at time *t+1* by playing ¬e. Likewise, E anticipates S_S's decision and, at time *t+1*, will revert to U_E in the PG. From the period *t+1* onwards, the payoffs of the game are determined by $(\neg e; U_E, D_{Ej})$. This case applies if $a + \lambda_E \geq d$.[16]

B) E defects at time *t* both in the PG (where it starts to play U_E) and in PD_{Ej} (where it plays D_{Ej}). In this case, the payoffs obtained by E and S_S in the PG at time *t*, are respectively *b* and *d*, which are determined by the strategy $(e,F;U_E)$. At time *t+1*, S_S will play ¬e because s/he believes that E will play U_E also at *t+1*. E

[15] Even though the structure of the PG is different from the PDs with regard to which we have defined the concepts of deficient and mutual relationship (section 2), by g_{EStkS} we mean the difference between the payoff obtained by E when it and Stk_S play F and the payoff that E obtains by defecting in the relation with Stk_S.

[16] If $a + \lambda_E < d$ it would be better for E to defect simultaneously in the PG and in the PD_{Ej}. See the following case B.

anticipates that S_S will not enter the PG at time *t+1* and continues to play U_E. For these reasons, from *t+1* onwards, we will observe in the repeated PG the strategies ($\neg e;U_E$) that generate the payoffs ($a+\lambda_{Ss}$, $a+\lambda_E$). This case applies if $a+\lambda_E < d$.

The discounted payoff obtained by E in the repeated PG when it and S_S repeatedly play fair by ($e,F;F_E$) is

$$(b+\lambda_E) / (1-\delta) \quad , \quad \text{with } 0 \leq \delta \leq 1.$$

The discounted payoff obtained by E if it adopts the strategy described in case A is

$$(b+\lambda_E) + (a+\lambda_E)\ \delta\ / (1-\delta)$$

Given that $a+\lambda_E < b+\lambda_E$, it follows that

$$[(b+\lambda_E)/(1-\delta)] - [(b+\lambda_E) + (a+\lambda_E)\ \delta\ / (1-\delta)] > 0$$

Hence, E prefers to play (F_E,C_{Ej}) instead of adopting the strategy described in case A.

With regard to case B, the discounted payoff obtained by E is

$$(b+\lambda_E) + d\delta + (a+\lambda_E)\ \delta^2 / (1-\delta).$$

Also in this case, given the assumption $b+\lambda_E > d$, it follows that

$$[(b+\lambda_E) / (1-\delta)] - [(b+\lambda_E) + d\delta + (a+\lambda_E)\ \delta^2 / (1-\delta)] > 0.$$

We conclude that, if E and S_S start to play ($e,F;F_E$), and if they reason as if S_S were endowed with high environmental cognitive social capital, and if E announces CSR principles that allow for the formation of reciprocal beliefs and conformist preferences (section 4), there are no incentives for E to stop playing F_E. This is true independently of the value of the discount factor δ.[17]

[17] In respect to the sub-game perfection of the 'fair equilibrium' in the PG, an alternative argument may be based on the demonstration (section 3) that (if λ is high enough as we assume in this case) S_S's threat

Since E does not have incentives to defect in the PG, the decision to deviate can only be the consequence of the strategy adopted in the PD_{Ej} played by the firm E with its weak stakeholders. This could be possible, and we will verify whether it is the case that E decides to defect in the two PD_{Ej} in which it is involved with weak stakeholders, even though it knows that this decision terminates cooperation also in the repeated PG.[18] For this reason, it is necessary to investigate the incentives that characterize E in the repeated PD_{Ej} with weak stakeholders. The stage-game normal form of the PD_{Ej} is shown in Figure 15.

Figure 15 The normal form PD_{Ej} stage game involving E and S_{Wj}

	C_{SwjE}	D_{SwjE}
$C_{E,Swj}$	*b, b*	*0, c*
$D_{E,Swj}$	*c, 0*	*a, a*

where $c > b > a > 0$.

The assumption is that, in repeated PD_{Ej}, player E's myopic value of δ does not make repeated cooperation sufficiently desirable for it. In other words, at any time *t*, the firm prefers to defect when the weak stakeholders play C_{SwjE} instead of continuing to cooperate, even though after the defection, the payoff that E obtains from the period *t+1* onwards is equal to $a\delta^{t-1}/(1-\delta)$. The deviation of E at the first stage represents the first opportunity for it to obtain the maximum advantage by defecting when S_{Wj} plays C_{SwjE}. In fact, given δ so that $g_{ESwj} < 0$, it follows that

$$[b/(1-\delta)] < \ldots < [b + b\delta + \ldots + b\delta^{t} + c\delta^{t+1} + a\delta^{t+2}/(1-\delta)] < [b + c\delta + a\delta^{2}/(1-\delta)] < [c + a\delta/(1-\delta)]$$

According to this payoff structure, if we consider only the PD_{Ej}, material incentives induce E to defect at the first stage in PD_{Ej} because at this first stage the incentive for E to defect (i.e the difference $g_{ESwj} < 0$) is the greatest. The payoff which E obtains by defecting at the first stage is

$$[c + a\delta/(1-\delta)]$$

---

of punishing the enterprise if it defects is a credible threat (see on this point Geanakoplos, Pearce and Stacchetti 1989).

[18] According to our definition, E prefers to defect with weak stakeholders in PDs instead of cooperating with them.

and

$$g_{ESwj} = [b / (1-\delta)] - [c + a\delta /(1-\delta)] < 0.$$

However, in order to explain E's behaviour, we must consider that E's decision to defect at the first stage in PD_{Ej} implies the sanction by S_S in the subsequent repeated PG games. In particular, according to the previous considerations, S_S will play ¬e in the PG the stage after E has defected in PD_{Ej}. By anticipating S_S's intention, E will stop playing F_E in the PG as well. For this reason, in order to understand the optimality of player E's strategy, we should simultaneously consider the payoff structure in the repeated PD_{Ej} and PG.

In particular, if

$$\sum_{j \in RE} g_{Ej} = g_{ESs} - (g_{ESw1} + g_{ESw2}) \geq 0$$

then we can demonstrate that if E, S_S and S_{Wj} start their relationship in a fair-cooperative way, there are no incentives for the firm to defect.

Consider the numerical example introduced in Figure 6 (section 3.4) and fix the following parameters $b=2$, $d=4$, $a=1$, $c=3$, $\lambda=3$ and $\delta_E = 0.41$. When E defects at the first stage in the PD_{Ej} and at the second stage in the PG, we obtain: $g_{ESs} = 0.695$, $g_{ESwj} = -0.305$ *and* $g_{ESs} - 2(g_{ESwj}) = 0.085$. This result holds independently of the stage when the firm may decide to defect.

For example, if E defects at the third stage in PD_{Ej} (and consequently, at the fourth stage in PG, the outcome is (¬e,U_E)), we obtain $g_{ESs} = 0.107$, $g_{ESwj} = -0.053$ and $g_{ESs} - 2(g'_{ESwj}) = 0.001$. For this reason, given these values of parameters and $\delta = 0.41$, the sub-network of E's cooperative relations is sustainable when players implement their MG trigger strategies. It is finally important to identify for which values of δ this result holds and E consequently prefers playing fairly and cooperatively in the repeated PG and PD_{Ei} respectively, instead of defecting.

First of all note that, when S_S and E are endowed by cognitive social capital (i.e. λ_E and λ_{Ss} are both $> d - b$ and only $\lambda_{Ss} > c - a$), the repeated-Fair F_E strategy in the PG is more profitable than the repeated-unfair U_E one independently of δ . In fact, every strategy of deviation induces the PG-stage-game equilibrium (¬e;U_E) and generates the repeated PG payoff $(a + \lambda_E)\delta^{t-1}/(1-\delta)$. This payoff is strictly lower than the psychological payoff of the repeated Fair strategies (consistent with playing the

modified MG trigger strategies defined in the preceding sections) inducing the PG-stage-game equilibrium (e,F;F_E) and the repeated PG payoff $(b + \lambda_E)\ \delta^{t-1} / (1-\delta)$. Thus, E will always cooperate in the PG.

Nevertheless, Fair play in the repeated PG becomes less and less profitable in comparison with defection when δ decreases while defection in PD_{Ej} becomes more and more profitable when δ decreases. For this reason, there will be a value δ^* which indicates the lowest value of player' E personal discount factor δ_E in correspondence to which it is still convenient for E to play fairly and cooperate (i.e. the stage-game strategy (F_E,C_{Ej})), while when δ_E is lower than δ^* E has incentives to defect in all three adjacent games. Thus, when $\delta_E < \delta^*$, E will defect in both the repeated PG and the two subsequent PD_{Ej} and the cooperative equilibrium will not be sustainable.

Using the previous numerical parameters ($b=2$, $d=4$, $a=1$, $c=3$, $\lambda=3$), we obtain δ^* equal to 0.4. In fact when $\delta_E = 0.4$, given the other values of parameters, it holds that $g^*_{ESs} - 2(g^*_{ESwj}) = 0$. For any value $\delta_E < \delta^*$ we have that the fair-cooperative repeated equilibrium fails. For example, if $\delta_E = 0.39$, $g_{ESs} = 0.6394$, $g_{ESw} = -0.3607$, *and* $g_{ESs} - 2(g_{ESwj}) = -0.082$.

The critical value δ^* can be calculated in general as a function of the parameters b, d, a, c, and λ_E of player E's payoff function when it compares the fair-cooperative iterated payoff and the payoff from the best deviation strategy. The relevant gains are respectively $g_{ESs} = [(b+\lambda_E)/(1-\delta)] - [(b+\lambda_E) + (a+\lambda_E)\ \delta/(1-\delta)]$ as far as the repeated PG is concerned, and $g_{ESwj} = [b/(1-\delta)] - [c + a\delta/(1-\delta)]$ in relation to the repeated PD_{Ej}

Note that the gain g_{ESs} can be simplified by $(b-a)\ \delta/(1-\delta)$, which, given the game's PG parameters, is in general a positive gain. Moreover, the gain g_{ESwj} can be simplified by $(b-c) + (b-a)\ \delta/(1-\delta)$,which due to the negative value of $(b-c)$ and the assumptions of the other parameters in this game is in general a negative gain. Thus in order to find δ^*, it must be established (recall that the negative gain from cooperation are doubled given that E plays two PD_{Wj}) that

$$(b-a)\ \delta^*/(1-\delta^*) = -2[(b-c)+ (b-a)\ \delta^*/(1-\delta^*)]$$

and that, given the negative value of the difference $(b-c)$ entails

$$(b-c) = -1/2\ [(b-a)\ \delta^*/(1-\delta^*)] - (b-a)\ \delta^*/(1-\delta^*) = 1.5[(b-a)\ \delta^*/(1-\delta^*)]$$

that is

$$\frac{(b-c)}{(b-a)}\ (-2/3) = \delta^* /(1-\delta^*)$$

In fact, according to our parameter $\delta^* = 0.40$, which is the solution for $(b-c)/(b-a) = -1$ (as it is in our case) and for $\delta^* /(1-\delta^*) = 0.666$.

We may conclude that the introduction of psychological payoffs into the game played between E and S_S – payoffs which stem from the agreement on the principle *T* (= CSR) of fairness in the pre-play communication phase of the game – makes the network among the firm and all its stakeholders sustainable, for values of player E's discount factor δ_E such that $\delta_E \leq \delta^*$, even though the firm has no material incentive to cooperating with weak stakeholders.

8 Conclusions

The aim of this chapter has been to investigate the theoretical relationship between social capital and corporate social responsibility. Our principal purpose has been to highlight the importance of cognitive social capital and CSR principles in generating cooperative networks between the firm and all its stakeholders (structural social capital).

Cognitive social capital consists of dispositions and beliefs functional to the development of conformist motivations that affect the agents' propensity to behave in different ways. Beliefs focus on reciprocal behaviours among agents and are affected by agreements on general principles and default reasoning stemming from agreements; but they also depend on the behaviour that other agents have exhibited in the past. Dispositions spring principally from the cultural environment of the most general social norms and values shared in society at large, so that they have a component independent of specific agreements on small-scale social norms and principles of behaviour, such as the CSR principle that a firm may agree with its stakeholders. But they also depend on micro elements (e.g. genetic and psychological factors) and cannot be activated without the other components of cognitive social capital that we have seen are related to more intentional elements like agreements on CSR norms. Conformist motivations are reasons to act in compliance with agreed principles of justice, such as CSR principles, and they are proportional to the level of conformity that an agent may reach through

his/her action contingently on his/her beliefs about other agents' behaviours, and they also depend on the expected reciprocity of other agents in obtaining high levels of conformity contingent on their own expectations about other agents. Conformist motivations operate as weights that determine the extent to which the exogenous and primitive cooperative dispositions can affect actual behaviors.

Structural social capital is understood as a global (multilateral) property of a relational network linking agents (for example firms and stakeholders) so that, independently of the deficiency of the specific bilateral relations, linkages in the network are nevertheless characterized by cooperation among agents. The sustainability of such linkages, and hence the possibility of observing a network structurally characterized by social capital, depends on four factors: a) reciprocal beliefs that others will cooperate, b) a generic disposition to cooperate, c) conformist motivations contingent on agreed norms and beliefs, d) the existence of sanctions against agents that decide not to cooperate. While the first three elements are cognitive components of social capital, the fourth is a structural characteristic of the game forms whereby interaction amongst agents takes place.

In this context, CSR is an essential part of the cognitive social capital that agents characterized as firms and stakeholders may possess to make cooperation in a relational network sustainable. In particular, CSR principles are the basis for impartial agreements amongst agents (firms and stakeholders) on which depend mutual beliefs concerning the level of principle compliance and conformist motivations (preferences) related to each of the solutions that agents can give to their interaction.

In regard to the firm's stakeholders, we have introduced a distinction between strong and weak stakeholders. The firm is interested in cooperating in the long term with strong stakeholders, and it is not interested in doing so with weak ones.

We have based our analytical framework on the relational network literature, and with particular regard to Lippert and Spagnolo (2009). But we have made an important innovation to this framework by introducing the idea of modelling at least some relations by means of psychological games. Thanks to this analytical model, we have been able to show that the agreement between the firms and its strong stakeholders on CSR fairness principles, which in their turn activate the other components of the firm's

and stakeholders' cognitive social capital, generates endogenous incentives for the firm to cooperate with weak stakeholders and creates cooperative relations that would otherwise not exist.

Our argument has consisted of the five following points:

1. In a context characterized by strong dispositions to conform with norms of fair cooperation (high levels of λ), and by the decision of the firm to agree with its strong stakeholders – belonging to the same context – on a contractarian principle of fair treatment addressed to whatever stakeholder (a principle of CSR), the effective implementation of such a social norm may stem from the fact that effective conformist preferences can be formed which activate the motivational force of cooperative dispositions. Thus individuals (both members of the organisation in a position of authority – the firm – or internal and external key stakeholders) will be induced by the motivational force of those dispositions to maintain fair and cooperative conduct also with respect to weak stakeholders. In other words, a CSR principle will be complied with even if there is no direct advantage in terms of material payoffs accruing to the powerful members of the organisation or to their strong stakeholders.

2. Dispositions do not operate in a vacuum. The agreement on a CSR principle may also favour the appropriate reciprocal beliefs concerning mutual conformity that by themselves furnish reasons to comply with the principle. The implementation of a CSR standard contributes to generating the belief in the firm's stakeholders that the firm will share cooperative relations with them. It is only with reference to explicit agreements on CSR principles that stakeholders can form their beliefs about the type of firm to which they are related.

3. This is a sort of moral reputation that reinforces cooperation which is not based only on the pursuit of material advantages. It therefore supplements the reasons for combining a good reputation with more intrinsic reasons to act.

4. The beliefs and dispositions related to cognitive social capital induce the strong stakeholders to cooperate with the firm if and only if it is also cooperative with weak stakeholders.

5. The possibility that strong stakeholders decide not to cooperate with the firm if it defects with weak stakeholders is a reliable threat for the firm, which may decide (it depends on the payoff structure) to cooperate with weak stakeholders in order to avoid the sanction from strong stakeholders.

6. This produces structural social capital (in terms of a sustainable network of cooperative relations involving the firm, the strong and the weak stakeholders) that would not be feasible without the threat of sanction by the strong stakeholders. This sanction is not due to exogenous reasons; rather, it is determined by endogenous incentives that we have explained by considering the effect of cognitive social capital on stakeholders' behaviour.

Our analysis has shown that there exists a Nash equilibrium which implies cooperation between the firm and all its stakeholders, both the strong and the weak ones. This cooperative equilibrium is sub-game perfect and it applies, for a reasonable value of the firm's discount factors δ, when the firm generates the appropriate belief in strong stakeholders – characterized by cognitive social capital in terms of disposition – by declaring a CSR standard.

Our findings raise numerous questions and ideas for further research.

First, they open the way to studies aimed at empirical verification of the effect of cognitive social capital and CSR declaration on cooperative behaviours by firms towards weak stakeholders.

Second, by shedding light on a new important role of SC, they encourage further theoretical and empirical analysis of the factors and the policies which may be able to increase cognitive social capital in terms of disposition to cooperate, which is a key element in fostering CSR adoption and cooperative relations between firms and weak stakeholders.

Appendix I

We report the formal representation of the function F which captures, for agents endowed by conformist preferences, the effects on ideal utility of beliefs in the degree

of conformity with the ideal by other agents (see also Grimalda and Sacconi 2002, 2005 Sacconi and Grimalda 2007, and Sacconi 2010c *infra*). We calculate the agents' ideal utility for each strategy pair of player S_S and the agents' ideal utility for E' strategies when it believes that S_S is going to play ¬e (in this respect note that E's ideal utility associated with its strategies - F_E and U_E - when E believes that S_S is going to play F or U may be easily computed by symmetrically considering the ideal utility of S_S when his/her strategies are (e,F) and (e,U) and his/her first-order beliefs are F_E or U_E).

The utility function of agents endowed with conformist preferences

The utility function of an agent i characterized by conformist preferences is:

$$V_i = U_i(\sigma) + \lambda_i F[T(\sigma)].$$

F is a function, shared by all the agents, of the normative fairness principle *T*. In abstract, *F* could be specified in different ways in order to consider various possible forms of the morality-grounded motive to behave, and it determines the weight of λ_i in the agents' gain. We follow Grimalda and Sacconi (2005) and Sacconi (2006) in adopting a particular specification for *F* based on an idea of expected mutuality in conforming with a contractarian principle of justice (*T*), captured by the Nash bargaining solution, which seems particularly coherent with the idea of an agreement involving the firm and its stakeholders (also called the Nash social welfare function N):

$$T(\sigma) = N(U_{1,\ldots,}U_n) = \prod_{i=1}^{n}(U_i - d_i)$$

where d_i stands for the reservation utility that player *i* can obtain when the bargaining process collapses. In the present context, we consider it appropriate to set all of these reservation utilities to zero.[19] To give an example related to the calculation of the value of *T*, consider the payoff matrix reported in Figures 4 and 5 (section 2.2.2), where the payoffs obtained by the three players – the firm, the strong and the weak stakeholder, i.e. the dummy player) are (2, 2, (2)) In this case, the principle *T* assumes

[19] This decision should be properly justified. Some authors argue that the proper choice for the "exit option" would be the Nash solution of the material game played in a non-cooperative way. However, this choice could be criticized because a possible situation of prevarication of one party over the other in the *status quo* would generate the final "moral" solution. For this reason, other authors have proposed the concept of a "moralised" status quo, where some minimal form of reciprocal respect is already in place.

the value $T = 2 \times 2 \times 2 = 8$. By contrast, when at least one player obtains a payoff equal to 0 (for example when the active players' strategies are (e,U;U_E), it is $T=3 \times 3 \times 0$).

Now, if we consider a two-person game, it is possible to define the two indices that contribute to determining F as follows:

1. $1 + f_i$: the index of player's i conditional conformity based on the degree of deviation from pure conditional conformity with T, that is, $f_i(\sigma_i, b_i^1)$:

$$f_i(\sigma_i, b_i^1) = \frac{T(\sigma_i, b_i^1) - T^{MAX}(b_i^1)}{T^{MAX}(b_i^1) - T^{MIN}(b_i^1)}$$

where $T^{MAX}(b_i^1)$ and $T^{MIN}(b_i^1)$ are the maximum and minimum values that the welfare distribution function, which represents the normative principle or ideology T, can assume, depending on i's action, given i's first-order belief , b_i^1, about the action that j is going to perform. $T(\sigma_i, b_i^1)$ is the actual level of T when player i carries out strategy σ_i given what s/he expects from player j. f_i varies from 0 (no deviation at all from the principle T) to -1 (maximal deviation).

2. $1+ \tilde{f}_j$: the index of player j's expected reciprocity in conformity based on the evaluation that player i forms about j's deviation from full conformity with the principle T , that is, $\tilde{f}_j(b_i^1, b_i^2)$:

$$\tilde{f}_j(b_i^1, b_i^2) = \frac{T(b_i^1, b_i^2) - T^{MAX}(b_i^2)}{T^{MAX}(b_i^2) - T^{MIN}(b_i^2)}$$

where b_i^1 is the first-order belief of player 1 about the action of player j. b_i^2 is the second-order belief about player j's belief in the action adopted by player i. $T^{MAX}(b_i^2)$ and $T^{MIN}(b_i^2)$ are the values that the welfare function takes when player j respectively maximises or minimises it, given the second-order belief of player i. In other words, $T^{MAX}(b_i^2)$ and $T^{MIN}(b_i^2)$ indicate the maximum and minimum value that player j can contribute to the welfare function, given his/her belief about i's action

---

Therefore, our choice (which follows Grimalda and Sacconi (2005) and Sacconi (2006)) may be considered equivalent to a notion of moralisation of the status quo from which the "bargaining" starts.

as perceived by *i* him/herself. $T(b_i^1, b_i^2)$ is the actual value that *i* expects the welfare function to take according to his/her beliefs. Also $\widetilde{f}_j$ varies between 0 and -1, which respectively indicate the maximum and minimum degree of conformity by player *j* with the ideology embodied in the welfare function *T*.

Implementing these definitions, the utility function of agent *i* can be written as:

$$V_i(\sigma_i, b_i^1, b_i^2) = U_i(\sigma_i, b_i^1) + \lambda_i[1 + \widetilde{f}_j(b_i^1, b_i^2)][1 + f_i(\sigma_i, b_i^1)]$$

Method for calculation of the agents' ideal utility

In this part we provide a detailed illustration of the method for calculation of the ideal utility component of the players' payoffs. The reference game and the parameters of the material part of the utility functions are those given in Figure 4 and Figure 5 (section 2.2.2) of the main text. The calculation complements the qualitative discussion conducted in section 3.3.

First, we must remember that the values of the agents' conformity indexes $[1 + f_i(\sigma_i, b_i^1)]$ and $[1 + \widetilde{f}_j(b_i^1, b_i^2)]$ result from the subtraction of a deviation measure ranging between 0 (no deviation at all from the principle) and -1 (complete deviation) from the unit (i.e. 1 means maximal conformity). Taking account of different possible belief systems (i.e. first- and second-order beliefs justifying the prediction of any given outcome of the game), the conformity indexes attached to how players carry out each state of the game may be computed.

- Strategy (e,F) of S_S given the first-order belief (b_{Ss}^1) that E plays F_E and given the second-order belief b_{Ss}^2 that E believes that S_S plays (e,F).

The deviation of player S_S from full conformity for strategy (e,F) is in this case:

$$f_{Ss}(e,F;F_E) = \frac{T(e,F;\, F_E) - T^{MAX}(F_E)}{T^{MAX}(F_E) - T^{MIN}(F_E)} = \frac{T(e,F;\, F_E) - T(e,F;\, F_E)}{T(e,F;\, F_E) - T(e,U;\, F_E)} = 0,$$

which entails a player S_S index of conditional conformity $[1 + f_{Ss}(e,F/F_E)] = 1$

Player E's expected deviation from full conformity for strategy F_E is in this case

$$\widetilde{f}_E(F_E;e,F) = \frac{T(F_E;\, e,F) - T^{MAX}(e,F)}{T^{MAX}(e,F) - T^{MIN}(e,F)} = \frac{T(F_E;\, e,F) - T(F;\, e,F)}{T(F;\, e,F) - T(U;\, e,F)} = 0,$$

so the *index* of expected reciprocal conformity is $[1 + \widetilde{f}_E\ (F_E; e,F)\ ] = 1$. Thus, in this case, player S_S's strategy (e,F) obtaining ideal utility is λ (recall that the ideal utility stems from $\lambda_i\,[1+\widetilde{f}_j\ (b_i^1, b_i^2)][1 + f_i\ (\sigma_i, b_i^1)]$)

- Strategy (e,F) of S_S, given the first-order belief (b_{Ss}^1) that E plays U_E and given the second-order belief b_{Ss}^2 that E believes that S_S plays (e,F).

The deviation of player S_S from full conformity for strategy (e,F) is in this case

$$f_{Ss}\,(e,F;U_E)\ \ \frac{T(e,F;\,U_E) - T^{MAX}\,(U_E)}{T^{MAX}\,(U_E) - T^{MIN}(U_E)} = \frac{T(e,F;U_E) - T\,(\neg e;\,U_E)}{T\,(\neg e;\,U_E) - T\,(e,U;\,U_E)} = -1,$$

which entails a player S_S index of conditional conformity $[1+ f_{Ss}\ (e;\,U_E)]=0$

Player E's expected deviation from full conformity for strategy U_E is in this cases

$$\widetilde{f}_E\,(U_E;\,e,F) = \frac{T(U_E;\,e,F) - T^{MAX}\,(e,F)}{T^{MAX}\,(e,F) - T^{MIN}(e,F)} = \frac{T(U_E;\,e,F) - T\,(F_E;\,e,F)}{T\,(F_E;\,e,F) - T\,(U_E;\,e,F)} = -1,$$

the index of expected reciprocal conformity is $[1 + \widetilde{f}_E\ (U_E;\,e,F)] = 0$. Thus, in this case, the ideal utility for player S_S's strategy (e,F) is 0.

- Strategy (e,U) of S_S, given the first-order belief (b_{Ss}^1) that E plays F_E and given the second order belief b_{Ss}^2 that E believes that S_S plays (e,U).

The deviation of player S_S from full conformity for strategy (e,U) is in this case

$$f_{Ss}\ (e,U;F_E) = \frac{T(e,U;\,F_E) - T^{MAX}\,(F_E)}{T^{MAX}\,(F_E) - T^{MIN}(F_E)} = \frac{T(e,U;\,F_E) - T\,(e,F;\,F_E)}{T\,(e,F;\,F_E) - T\,(e,U;\,F_E)} = -1,$$

player S_S's index of conditional conformity is therefore $[1 + f_{Ss}\ (e,U;\,F_E)] = 0$

Player E's expected deviation from full conformity for strategy F_E is in this case

$$\widetilde{f}_E\ (F_E;e,U) = T(F_E/e,U) - T^{MAX}\,(e,U) = 0$$

which entails an index of expected reciprocal conformity $[1 + \widetilde{f}_E\ (F_E;e,U)] = 1$. Thus the ideal utility in this case for player Ss's strategy (e,U) is 0.

The calculation of the expected reciprocal conformity index $[1+\widetilde{f}_E(F_E;e,U)]$ highlights a distinctive feature of conformity indexes in games such as the one considered in this chapter. When the strong stakeholder S_S believes that the other player E believes that s/he is going to play U, the maximum and the minimum value of the

function *T* (that may be generated by whatever response of player E to the strategy U) coincide. In these cases, the welfare distribution function, which represents the normative principle *T*, always takes value 0. This means that when the second-order belief of player Ss is U (that is, S_S believes that E believes that s/he is choosing U), s/he also believes that E cannot do any better by its choice than accept that the weak stakeholder will get 0. Thus, in these cases, a player - for example E - has no role in affecting the implementation of the principle *T*.

Note that if the maximum and minimum values of *T* are the same, the two differences at the numerator and the denominator in the deviation index are both 0, and the index is indefinite (you cannot divide by 0). However, since the only value admitted for *T* at the numerator is constant (so that also the difference at numerator is 0), it does not make sense to normalize the deviation from conformity in the interval from a maximum and a minimum value. In fact no deviation at all is allowed. Consequently, we will assume that in all cases like this (in particular note that the same reasoning applies when the second-order belief of Ss is (¬e)), the value of the expected reciprocal conformity index is the difference between the value of *T* determined by considering simply *the absolute value* of the difference between the (expected) choice *F* given the second-order belief that *(e,U)* is chosen (i.e $T(F_E; e,U)$) and the maximum value that *T* can take, again given the second-order belief that (e,U) is chosen (i.e. $T^{MAX}(e,U)$) (that is, what would be the numerator of the fraction normally representing the expected deviation from full reciprocal conformity).

- Strategy (e,U) of S_S, given the first-order belief (b^1_{Ss}) that E plays U_E and given the second-order belief b^2_{Ss} that E believes that S_S plays (e,U).

The deviation of player S_S from full conformity for strategy (e,U) is in this case

$$f_{Ss}(e,U;U_E) = \frac{T(e,U;\,U_E) - T^{MAX}(U_E)}{T^{MAX}(U_E) - T^{MIN}(U_E)} = \frac{T(e,U;\,U_E) - T(\neg e;\,U_E)}{T(\neg e;\,U_E) - T(e,U;\,U_E)} = -1,$$

which means that player S_S's index of conditional conformity is $[1 + f_{Ss}(e,U;U_E)] = 0$.

Player E's expected deviation from full conformity for strategy U_E in this case is similar to the previous case, and hence the same method of calculation applies.

$$\widetilde{f}_E\ (U_E;\ e,U) = T(U_E;\ e,U) - T^{MAX}\ (e,U) = 0$$

so that the index of expected reciprocal conformity is $[1 + \widetilde{f}_E\ (U_E;\ e,U)] = 1$. Again, the ideal utility of player SS for the strategy (e,U) under these contingencies is 0

- Strategy (¬e) of S_S, given the first-order belief (b^1_{Ss}) that E plays U_E and given the second-order belief b^2_{Ss} that E believes that S_S plays (¬e).

The deviation of player S_S from full conformity for strategy (¬e) is in this case

$$f_{Ss}\ (\neg e;U_E) = \frac{T(\neg e;\ U_E) - T^{MAX}\ (U_E}{T^{MAX}\ (U_E) - T^{MIN}(U_E)} = \frac{T(\neg e;\ U_E) - T\ (\neg e;\ U_E)}{T(\neg e;\ U_E) - T\ (e,U;\ U_E)} = 0,$$

which entails an index of conditional conformity of player S_S $[1 + f_{Ss}\ (\neg e;\ U_E)] = 1$

Player E's expected deviation from full conformity for strategy U_E in this case is similar to the previous case and hence the same method of calculation applies

$$\widetilde{f}_E\ (U_E;\ \neg e) = T(U_E;\ \neg e) - T^{MAX}\ (\neg e) = 0$$

which entails $[1 + \widetilde{f}_E\ (U_E;\ \neg e)] = 1$.

These two indexes of conditional and expected conformity jointly imply an ideal utility λ for the strategy (¬e) of player S_S under this case.

- Strategy (¬e) of S_S, given the first-order belief (b^1_{Ss}) that E plays F_E and given the second-order belief b^2_{Ss} that E believes that S_S plays (¬e).

The deviation of player S_S from full conformity for strategy (¬e) in this case is

$$f_{Ss}\ (\neg e;F) = \frac{T(\neg e;\ F_E) - T^{MAX}\ (F_E)}{T^{MAX}\ (F_E) - T^{MIN}(F_E)} = \frac{T(\neg e;\ F) - T\ (e,F;\ F)}{T\ (e,F;\ F_E) - T\ (e,U;\ F_E)} = -7/8$$

the index of conditional conformity of player S_S in this case is $[1 + f_{Ss}\ (\neg e\ |F_E)] = 1/8$

The expected deviation of player E from full conformity belongs to the class of cases (see also the discussion of the following case) that allow simple use of the absolute

difference between the *T* value for the expected choice of player E given the second-order belief about player S_S's choice $\neg e$ and the maximum value of *T* given $\neg e$

$$\widetilde{f}_E\ (F_E; \neg e) = T\ (F_E;\ \neg e) - T^{MAX}\ (\neg e) = 0$$

so that the expected index of player E's expected reciprocal conformity is $[1 + \widetilde{f}_E\ (F_E;\ \neg e)] = 1$. Thus the two indexes jointly imply an ideal utility equal to $1/8\,\lambda$.

Let us consider E's strategies when it believes that S_S is going to play ¬e.

- Strategy (F_E) of E, given the first-order belief (b_E^1) that S_S plays ¬e and given the second-order belief b_E^2 that S_S believes that E plays (F_E).

The deviation of player E from full conformity with the strategy (F_E) given ¬e cannot be but nil since this is a case where the maximum and minimum values of *T*, given player S_S's choice ¬e, are identical. Thus

$$f_E\ (F_E; \neg e) =\ T(F; \neg e) - T^{MAX}\ (\neg e) = 0$$

so the conditional conformity index of player E in this case is $[1 + f_E\ (F_E; \neg e)\ ] = 1$.

The strategy ¬e (and the first-order belief that Ss is going to implement that strategy) highlights the second distinctive feature of conformity indexes in the type of game we are considering. In this case the peculiarity depends on the fact that player Ss's strategy ¬e assigns the game the same result regardless of the other player's behavior, since it amounts to simply preventing interaction from occurring by a unilateral decision to stay out of it. When the strong stakeholder plays ¬e, it always generates the payoffs (1,1,1). Thus, in this case, the firm has no role in affecting implementation of the principle *T* (the value that the welfare distribution function, which represents the normative principle *T*, assumes is always 1 no matter what player E's choice is).

In other words, given the strong stakeholder's strategy ¬e , the firm E cannot do any better than accept the *T* value equal to 1 determined by player S_S's choice, which is the only one possible, and hence also the one with null deviation from the maximum value *T* possible when player S_S does ¬e. Also in this case, given that the E's first-order belief about player Ss's behavior is ¬e, as in the case discussed above, the general form of the conformity indexes would be indeterminate (the denominator of the fraction is 0), and again there can be only one constant value of *T* (at the numerator) . Therefore, in this case too, it does not make sense to normalize the deviation from conformity with

respect to the interval between maximum and minimum values of T, since no deviation is allowed at all. As we assume in all the cases like the one considered here, the deviation measure from the maximum possible value of *T* will be taken to be the simple absolute difference between the value of *T* determined as a consequence of player E's choice (given the ¬e choice of player S_S) and the maximum value of *T* possible under that choice (that is, the numerator of the fraction would typically represent the deviation from full conditional conformity).

Player S_S's expected deviation from full reciprocal conformity for strategy **¬e** is in this case an intermediate value

$$\widetilde{f}_{Ss}(\neg e; F_E) = \frac{T(\neg e; F_E) - T^{MAX}(F_E)}{T^{MAX}(F_E) - T^{MIN}(F_E)} = \frac{T(\neg e; F_E) - T(F; F_E)}{T(F; F_E) - T(U; F_E)} = -7/8$$

so that the index of expected reciprocity in conformity for the strategy $\neg e$ of player S_S is $[1 + \widetilde{f}_{Ss}(\neg e; F_E)] = 1/8$, which together with the aforementioned index of player E's conditional conformity gives to player E's strategy F given $\neg e$ the ideal utility $1/8\,\lambda$

- Strategy (U_E) of E given the first-order belief (b_E^1) that S_S plays ¬e and given the second-order belief b_E^2 that S_S believes that E plays (U_E).

The deviation of player E from full conformity by using strategy (U_E) given that S_S does $\neg e$ cannot be positive. Once again we have a case where, given the strategy choice of S_S player E cannot do any better than simply observe the decision of player S_S prevents the interaction from occurring and assigns a unique *T* value to the game, which, whatever player E's choice may be, cannot be different from $T = 1$,

$$f_E(U_E; \neg e) = T(U_E; \neg e) - T^{MAX}(\neg e) = 0$$

which entails for player E a conditional conformity index $[1 + f_E(U_E; \neg e)] = 1$

Finally, consider the expected deviation of player S_S from full reciprocity in conformity when s/he is believed to choose $\neg e$ given U_E.

$$\widetilde{f}_{Ss}(\neg e; U_E) = \frac{T(\neg e; U_E) - T^{MAX}(U_E)}{T^{MAX}(U_E) - T^{MIN}(U_E)} = \frac{T(\neg e; U_E) - T(\neg e; U_E)}{T(\neg e; U_E) - T(U; U_E)} = 0$$

the index of conditional conformity of player S_S is thus $[1 + \widetilde{f}_{Ss}(\neg e; U_E)] = 1$. Therefore when player E chooses F_E given S_S staying out, and E predicts that S_S does $\neg e$ jointly

the two indexes of conformity are fully positive and thus the ideal utility for player E is λ.

This concludes the calculation of the ideal utilities of players E and S_S for the different states of the PG game under the hypothesis that the players have mutually consistent beliefs systems about the game's outcomes.

References

M. Aoki (2010) *infra* 'Linking Economic and Social-Exchange Games: From the Community Norm to CSR', in L. Sacconi and G. Degli Antoni (eds), *Social Capital, Corporate Social Responsibility, Economic Behavior and Performance* (Basingstoke: Palgrave MacMillan).

K. Binmore (2005) *Natural Justice* (Oxford: Oxford University Press).

R. Burt (2002) 'The Social Capital of Structural Holes', in M.F. Guillen, R. Collins, P. England and M. Meyer (eds) *The New Economic Sociology* (New York: Russell Sage Foundation).

Clarkson Centre for Business Ethics (2002) 'Principles of Stakeholder Management', *Business Ethics Quarterly*, 12(2), 257–264.

J.S. Coleman (1988) 'Social Capital in the Creation of Human Capital', *American Journal of Sociology*, 94, 95–120.

R. Flannigan (1989) 'The Fiduciary Obligation', *Oxford Journal of Legal Studies* 9, 285–294.

E. Freeman (1984) *Strategic Management, A Stakeholder Approach* (Boston: Pitman).

E. Freeman (2000), 'Business Ethics at the Millennium', *Business Ethics Quarterly*, 10(1), 169–180.

T. Freeman and W.M. Evan (1990) 'Corporate Governance: A Stakeholder Interpretation', *The Journal of Behavioral Economics*, 19(4), 337–359.

R.E. Freeman and J. McVea (2002) 'A Stakeholder Approach to Strategic Management', Working paper n.01-02, Darden Graduate School of Business Administration.

M. Friedman (1977) 'The Social Responsibility of Business Is to Make Profits', in G.A. Steiner and J.F. Steiner, *Issues in Business and Society* (New York: Random House).

J. Geanakoplos, D. Pearce and E. Stacchetti (1989) 'Psychological Games and Sequential Rationality', *Games and Economic Behavior*, 1, 60–79.

G. Grimalda and L. Sacconi (2002) 'The Constitution of the Nonprofit Enterprise: Ideals, Conformism and Reciprocity', Liuc Papers n. 115, Serie Etica, Diritto ed Economia.

G. Grimalda and L. Sacconi (2005) 'The Constitution of the Not-For Profit Organization: Reciprocal Conformity', *Constitutional Political Economy*, 16(3), 249–276.

G. Grimalda and L. Sacconi (2005) 'Ideals, conformism and reciprocity: A model of Individual Choice with Conformist Motivations, and an Application to the Not-for-Profit Case', in P. L. Porta and L. Bruni (eds) *Handbook of Happiness in Economics* (Cheltenham Northampton, Mass.: Elgar).

S. Grossman and O. Hart (1986), 'The Costs and Benefit of Ownership: A Theory of Vertical and Lateral Integration', *Journal of Political Economy*, 94, 691–719.

H. Hansmann (1996) *The Ownership of Enterprise* (Cambridge, Mass.: Harvard University Press).

O. Hart (1995) *Firms, Contract and Financial Structure* (Oxford: Clarendon Press).

O. Hart and J. Moore (1990) 'Property Rights and the Nature of the Firm', *Journal of Political Economy*, 98, 1119–1158.

M.C. Jensen (2001) 'Value Maximization, Stakeholder Theory, and the Corporate Objective Function', *Journal of Applied Corporate Finance*, 14(3), 8–21.

S. Knack and P. Keefer (1997) 'Does Social Capital Have An Economic Payoff? A Cross Country Investigation', *The Quarterly Journal of Economics*, CXII, 1251–1287.

D. Kreps (1990) *Game Theory and Economic Modeling* (Oxford: Oxford University Press).

N. Lin (2001) *Social Capital* (Cambridge: Cambridge University Press).

S. Lippert, and G. Spagnolo (2009) 'Networks of Relations and Word-of-Mouth Communication', SSE/EFI Working Paper in Economics and Finance No 570, http://swopec.hhs.se.
S. Lippert (2010) *infra* 'Social Capital in Networks of Relations', in L. Sacconi and G. Degli Antoni (eds), *Social Capital, Corporate Social Responsibility, Economic Behavior and Performance* (Basingstoke: Palgrave MacMillan).
D. Narayan (1999) 'Bonds and Bridges: Social Capital and Poverty', *Poverty Group* PREM, The World Bank.
J. Nash (1950) 'The Bargaining Problem', *Econometrica*, 18, 155–162.
M. Paldam (2000), 'Social Capital: One or Many? Definition and Measurament', *Journal of Economic Surveys*, 14(5), 629–653.
R.D. Putnam, R. Leonardi and R.Y. Nanetti (1993) *Making Democracy Work: Civic Traditions in Modern Italy* (Princeton: Princeton University Press).
M. Rabin (1993) 'Incorporating Fairness into Game Theory', *American Economic Review*, **83** (5), pp. 1281–1302.
L. Sacconi (1999) 'Codes of Ethics As Contractarian Constraint on Abuse of Authority: A Perspective from the Theory of the Firm', *Journal of Business Ethics*, 21, 189–202.
L. Sacconi (2000) *The Social Contract of the Firm, Economics, Ethics and Organisations* (Berlin: Springer Verlag).
L. Sacconi (2004) 'Incomplete Contracts and Corporate Ethics: A Game Theoretical Model under Fuzzy Information', in F. Cafaggi, A. Nicita and U. Pagano (eds) *Legal Orderings and Economic Institutions* (London: Routledge).
L. Sacconi (2006) 'A Social Contract Account For CSR as Extended Model of Corporate Governance (I): Rational Bargaining and Justification', *Journal of Business Ethics*, Special Issue on Social Contract Theories in Business Ethics, 259–281.
L. Sacconi (2007a), 'A Social Contract Account for CSR as Extended Model of Corporate Governance (II): Compliance, Reputation and Reciprocity', *Journal of Business Ethics*, 75(1), 77–96.
L. Sacconi (2007b), 'CSR as a model of extended corporate governance, an explanation based on the economic theories of social contract, reputation and reciprocal conformism', in F. Cafaggi (ed.) *Profiles of Self-Regulation* (Boston, Mass: Kluwer Academic Press).
L. Sacconi (2010a) 'A Rawlsian view of CSR and the Game Theory of its Implementation (Part I): The Multistakeholder Model of Corporate Governance', in L. Sacconi, M. Blair, E. Freeman and A. Vercelli (eds) '*Corporate Social Responsibility and Corporate Governance: The Contribution of Economic Theory and Related Disciplines*', Basingstoke: Palgrave Macmillan, in print.
L. Sacconi (2010b) 'A Rawlsian view of CSR and the Game Theory of its Implementation (Part II): Fairness and Equilibrium', in L. Sacconi, M. Blair, E. Freeman and A. Vercelli (eds) '*Corporate Social Responsibility and Corporate Governance: The Contribution of Economic Theory and Related Disciplines*', Basingstoke: Palgrave Macmillan, in print.
L. Sacconi (2010c) *infra* 'A Rawlsian View of CSR and the Game Theory of its Implementation (III): Conformism and Equilibrium Selection', in L. Sacconi and G. Degli Antoni (eds), *Social Capital, Corporate Social Responsibility, Economic Behavior and Performance* (Basingstoke: Palgrave MacMillan).
L. Sacconi and G. Degli Antoni (2009) 'A Theoretical Analysis of the Relationship between Social Capital and Corporate Social Responsibility: Concepts and Definitions', in S. Sacchetti and R. Sugden (eds) *Knowledge in the Development of Economies. Institutional Choices under Globalisation* (Cheltenham: Edward Elgar), pp. 134–157.
L.Sacconi and M. Faillo (2010) 'Conformity, reciprocity and the sense of justice. How social contract-based preferences and beliefs explain norm compliance: the experimental evidence' *Constitutional Political Economy*, 21(2) 171-201.
L. Sacconi, S. DeColle and E. Baldin (2003) 'The Q-RES Project: the Quality of Social and Ethical Responsibility of Corporations', in J. Wieland (ed.) *Standards and Audits for Ethics Management Systems, The European Perspective* (Berlin: Springer Verlag), pp. 60–117.

R. Selten (1967): 'Die Strategiemethode zur Erforschung des eingeschriinkt rationalen Verhaltens im Rahmen eines Oligopolexperiments' in H.. Sauermann (ed), *Beiträge zur Experimentellen Wirtschaftsforschung*, Vol.1 Tübingen: J.C.B. Mohr (siebeck), 136-168.

R. Selten (1975) 'Reexamination of the Perfectness Concept for Equilibrium Points in Extensive Games', *International Journal of Game Theory*, 4, 25–55.

N. Uphoff (1999) 'Understanding Social Capital: Learning from the Analysis and Experience of Participation', in P. Dasgupta and I. Serageldin (eds) *Social Capital: A Multifaceted Perspective* (Washington, DC: The World Bank), pp. 215–249.

O. Williamson (1975) *Market and Hierarchies* (New York: The Free Press).

O. Williamson (1986) *The Economic Institution of Capitalism* (New York: The Free Press).